Modeling the Decision Making Mind: Does Form Follow Function?

Dissertation

zur Erlangung des akademischen Grades
Doktor der Psychologie

eingereicht an der

Lebenswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

von

Dipl.-Psych. Jana Jarecki, B.Sc. Econ

Präsidentin der Humboldt-Universität zu Berlin: Professorin Dr.-Ing. Dr. Sabine Kunst

Dekan der Lebenswissenschaftlichen Fakultät: Professor Dr. Bernhard Grimm

Gutachter

I. Professor Gerd Gigerenzer
II. Professorin Elke van der Meer
III. Professor Wolfgang Gaissmaier

Tag der Verteidigung: 08. Juni 2016

Acknowledgements

First and foremost, thanks to my parents,Wolfgang and Inge. Wolfgang, you are the most caring, understanding father one can wish for. I inherited (or learned) the love for mathematics, and logics, and thinking out of the box from you. Inge, I have gotten the love for psychology from you. I am sad that you are not with me anymore, I know you would be proud. Further, many thanks to my brothers Julian and Jeffi for taking over when I needed you to. I love you.

Two parallel roads led me toward this PHD. The first was encountering Professor Norbert Bischof in 2010, who was the first researcher to introduce me to cognitive modeling, back in Munich, and it progressed with meeting Michael Zehetleitner. The second road concerns Gerd Gigerenzer and the ABC research group, whom I encountered while searching for a place that merges economic approaches, cognitive modeling, and basic psychology. It took me half a year to get the courage to apply for an internship at this group (thanks to David Bauder) for kicking my butt to apply! :)). I am still impressed how Gerd makes this group work (it may be the cake ...).

In 2010 I met my then-to-be supervisor, Jonathan Nelson. Without your encouragement, support, and advise I would not be where I am right now. Thanks for all your guidance, not only regarding the content of science, but also regarding presentation skills, and shaping me as scientific persona. What I am especially glad for is that your support was steady throughout all of these three years. Much the same gratitude extends towards my second main supervisor, Björn Meder. Thank you so much for your guidance, the structure provided, and also for really supporting me wherever possible. I have learned a tremendous amount from your "top level" and "fine level" comments. Looking back, the feature which maybe had the greatest impact on my development is that the both of you were unconditionally supportive of my crazy and sane ideas, all new projects, and that you encouraged me to go the one step beyond what I thought was possible. During my time at the MPI I met Jolene Tan with whom I shared a passion for meta-theories, and evolutionary psychology. Jolene — thank you for being an amazing sparring partner for thought exchange on process modeling and also evolutionary psychology. I enjoyed it so much to run my ideas by you and will never forget our digital exchange of arguments (highly recommended: scientific discussions in written form!). Early in 2013 I met Mirjam Jenny. Our first encounter concerned the joint organization of the Summer Institute 2013. I found you immediately inspiring! Thanks for our discussions about models and modeling, and moreover for getting the process tracing community in touch! And finally, also in 2013 I met Andreas Wilke when he gave a talk at a conference somewhere in the US about evolutionary approaches to risky choice. Thanks for your input, thoughts, support, and enthusiasm!

Thanks to Anita Todd and (in no particular order) Nathaniel, Jonathan, David, Hendrik, Björn, Eiko, and Christian, for helpful comments and proofreading. Thanks to Gregor for the data collection.

Thanks also to the Max Planck Society and the IMPRS uncertainty for their financial support and

encouragement.

Last but not least there are my friends, old and new. Where would I be without you? JOLENE, your scientific qualities are totally matched by your great heart, thanks for being the best friend alive (and for feeding me ;)). HANNA, I am glad I got to know you and for the motivation and empathy! Thanks to JACOB for all of our fundamental discussions about assumptions in science, I have learned from you. A warm thanks also to CHRISTIAN for all the smiles you made me smile. Especially I want to thank NIKLAS for the roof over my head!

Abstract

The behavioral sciences take two distinct but complementary approaches toward explaining decision making: the form of information processing and the function of the resulting behavior. Or *how* the mind works and for which *functions* it works. Cognitive scientists mostly study fine-grained cognitive processes, while behavioral economists, evolutionary psychologists, behavioral biologists, and evolutionary anthropologists traditionally study the performance and functions of decision behavior. My dissertation argues for form-function integration.

Oftentimes, functional analyses of decision making involve optimization models. A crucial step toward form-function-integration is supplementing the optimality-type models of evolutionary psychology and biology (and part of psychology) with cognitive process models. In study 1 I propose a general conceptual framework to guide the development of process models. Results of a literature review and a survey among researchers established that behavioral scientists rely increasingly on cognitive modeling, and moreover that the term 'process model' is widely-used, but at the same time ill-defined. My novel framework for process models conceptualizes them as models that contain at least one intermediate stage, make testable predictions for both the cognitive process and the resulting final decision, and are built such that the process prediction can be derived from the models' input separately from the behavior, which is derived from input plus process. This avoids reverse inference (erroneously inferring the truth of a state from the truth of its consequence). Further the computations of process models need to be compatible with the current state of knowledge (i.e., it needs to be plausible).

A different but complementary approach to form-function integration is to analyze decisions from a computationally functional stance. This concerns solving complex challenges posed by the environment efficiently. In study 2 I argue that *robustness* is a key property needed for the human categorization system to perform functionally. This is due to two computational challenges inherent in classification. The first challenge is a combinatorial explosion: as the number of features grows, the number of feature-category combinations grows exponentially. The second challenge concerns limited experience: often there are only a few previous instances available from which the cognitive system can generalize to classify a novel exemplar. Robustness – an important property within system's biology and the framework of ecological rationality – refers to the persistence of function across different environments (in evolutionary biology it also refers to limits to evolvability). Regarding categorization, there is one categorization algorithm from machine learning that has been found to be particularly robust against violations of its underlying assumption: Naïve Bayes. The algorithm assumes that the features that make up exemplars are statistically independent given the true class of the exemplar. This is a simple yet robust principle. My study combined computational cognitive modeling and simulations with data from two categorization learning experiments to study whether people behave according to class-conditional independence, specifically whether they use it as default when they begin categorizing. Results from two experiments show that human errors early in learning, the aggregate learning curves, and the individual behavior are best described by adhering to class-conditional independence.

A proper bridge between form and function involves bridging levels of analysis. An often discussed class of functions refers to coping with adaptive problems. Evolutionary psychology assumes that modern human cognitive processes are, in part, recruited from a set of processes tailored to cope with *survival and reproductive challenges* in the past. In study 3 I argue that human risk taking and the underlying processes serve evolutionary functions. I argue that most behavioral discrete choice experiments on risk erroneously assume that risk-taking is a unitary, general cognitive phenomenon independent of the goals of the decision maker. On the other hand, many questionnaires find that peoples' propensity to take risks depends on the domain, but this work makes few explicit cognitive processing assumptions. Questionnaires measure behavioral tendencies, but not the cognitive processes. I investigated the functions of risk as depending on evolutionary domain. I found that there are stable domain differences across ten different domains. Furthermore, while previous research has found that in general women are less risk taking than men, my results show that, although the general claim is true, women are in some domains more risk taking than men. Further, I investigated the cognitive processes underlying these domain-specific risk propensities by investigating the situational aspects related to risk taking (cues) that people retrieve and the cue direction (i.e., whether the cue favors or disfavors engaging in the risk). While previous studies conducted in the financial domain suggested that the order of positive and negative reasons determines choice outcomes, the results show that neither the frequency of the cue direction, nor the order of the cue direction are related to differences in risk propensities across domains. Rather, the frequencies of specific cues are related to risk taking.

Zusammenfassung

Die Verhaltenswissenschaften betrachten menschliches Entscheiden aus zwei komplementären Blickwinkeln: von Seiten der Form und der Funktion. Formfragen behandeln wie *Denkprozesse* ablaufen, wohingegen Funktionsfragen sich damit befassen welches *Ziel* das resultierende Verhalten erfüllt. Psychologische Kognitionsforschung untersucht in detaillierten fomalen Modellen welche Prozesse den Entscheidungen zu Grunde liegen. Verhaltensökonomie, evolutionäre Psychologie, Verhaltensbiologie, und evolutionäre Anthropologie befasst sich damit, welche Performanzkriterien die resultierenden Entscheidungen erfüllen. Die vorliegende Dissertation argumentiert füür die Integration von Form und Funktion.

Ein Schritt zur Integration von Form und Funktion besteh darin Prozessmodelle aus der Kognitionsychologie in die evolutionäre Psychologie und Verhaltensbiologie einzuführen. Studie 1 untersucht die Eigenschaften kognitiver Prozessmodelle. Da die Kognitionsforschung nicht klar spezifiziert, welche formalen oder inhaltsbezogenen Eigenschaften ein Modell als Prozessmodell qualifizieren, schlage ich ein Rahmenmodell für allgemeine kognitive Prozessmodelle vor, mit Hilfe dessen Prozessmodelle entwickelt werden können. Das Rahmenmodell besagt dass Prozessmodelle sich durch folgende Eigenschaften auszeichnen: Sie spezifizieren zumindest einen Prozesschritt explizit, erlauben testbare Vorhersagen für sowohl den Prozess als auch das resultierende Verhalten, ihre Struktur erlaubt es die Prozessvorhersagen aus den Eingangsgrößen abzuleiten und die Verhaltensvorhersage aus Eingangs- und Prozessgrößen abzuleiten, statt von guten Verhaltensmodellierungen auf Prozesse zu schließen. Zudem sollten die kognitiven Operationen innerhalb eines Prozessmodells mit dem bestehenden Wissen über kognitive Kapazitäten übereinstimmen (Plausibilität).

In Studie 2 untersuchen wir Kategorisierungsentscheidungen aus Perspektive der Form und Funktion. Es geht hierbei um robuste Performanz des menschlichen Kategorisierungssystems. Wie *robust* ein System gegenüber Umweltveränderungen reagiert, stellt eine der zentralen Frage der Biologie sowie der ökologischen Rationalität dar. Ein Kategorisierungsalgorithmus aus der Informatik zeichnet sich durch seine Robustheit aus: Naïve Bayes. Der Algorithmus implementiert die statistische Annahme, dass die Merkmale eines Objektes statistisch unabhängig gegeben die wahre Klasse des Objektes sind (klassenkonditionale Unabhängigkeit). In vielen Situationen in denen die Merkmale der objekte *nicht* unabhängig sind, wählt der Algorithmus nichtsdestotrotz die richtige Klasse. Wir untersuchten ob Menschen beim Klassifizieren implizit von klassenkonditionaler Unabhängigkeit ausgehen und ob sie diese Annahme im Laufe des Lernprozesses ablegen. Dazu implementierten wir klassenkonditionale Unabhängigkeit in einem Lernmodell und testeten die resultierenden Vorhersagen anhand von menschlichen Entscheidungsdaten in zwei computerbasierten Klassifikationsexperimenten. Wir finden, dass die Fehlerraten, die aggregierte Lernkurven, und die individuellen Entscheidungen mit der Annahme übereinstimmen, dass Menschen am Anfang des Lernprozesses gemäß des robusten Prinzips der klassenkonditionalen Unabhängigkeit Kategorisierungsentscheidungen treffen.

In Studie 3 geht es um Risikoentscheidungen aus der Perspektive der Form und Funktion. Sind Informationsverarbeitungsprozesse abhängig von der Zielgröße der Entscheidung? Wir untersuchen, inwieweit Menschen bei Entscheidungen über Handlungen deren Ausgang ungewiss ist in verschiedenen Lebensbereichen verschiedene Informationen verarbeiten. Die Risikoliteratur teilt sich in zwei Lager. Experimentelle Arbeiten zur menschlichen Informationsverarbeitung über Risiken präsentieren Menschen überwiegend mit finanziellen Lotterien. Dies impliziert dass monetäre Kosten-Nutzen-Überlegungen repräsentativ seien für die Fülle an Risiken und Unsicherheiten über die Menschen entscheiden. Arbeiten zur Verschiedenartigkeit menschlichen Risikoverhaltens in verschiedenen Lebensbereichen verwenden zumeist Fragebögen um Risikoeinstellungen zu messen. Fragebögen erfassen Verhaltenstendenzen, jedoch nicht die zugrundeliegenden Prozesse. Unsere Studie misst Prozess- und Verhaltensvariablen. Wir erfassen entscheidungsrelevante Information in zehn Domänen mittels offener Fragen und untersuchen wie sie mit Risikoeinstellung über Inhaltsdomänen hinweg zusammenhängt. Die Inhaltsdomänen sind hierbei evolutionär motiviert. Wir finden (a) dass sich Einstellungen zu Risiken systematisch über die Domänen hinweg unterscheidet. Insbesondere finden wir, dass Frauen nicht universell weniger Risiken als Männer eingehen, sondern dass Geschlecht und Funktion die relativen Risikopräferenzen bestimmt. Außerdem finden wir (b) dass weder die Anzahl der Einflüsse die positiv gerichtet sind, noch die Reihung der positiven und negativen Einflüsse, aber die spezifischen Situationsvariablen die unterschiedlichen Risikoeinstellungen über die Domänen hinweg erklären.

Contents

Chapter 1

Introduction: Form and function

In 350 BCE, Aristotle considered formal aspects (like the essential shape of a body) and final aspects (such as the health of a person) as two of his four constituents of any explanation (Aristotle, 'Physics', 350 BCE, trans. 2016). During behaviorism, Watson (1913) deemed external stimuli, and Tolman (1925) deemed goals to be the causes of behavior. Modern computational cognitive science separates Marr's (1982) computational 'why'–level from the algorithmic 'how'–level of analysis (see also Tinbergen, 1963). Theoretically, form-centered and function-centered research should converge in their findings and methodologies. Yet, most authors study either form or function with different methods (as will be outlined below), except for the study of the ecological rationality of heuristics (Todd, Gigerenzer, & the ABC Research Group, 2012).

In my dissertation I wish to bridge form and function. I first propose a general approach to this; namely cognitive processes evolved to cope with evolutionary problems. I argue:

— Discovery of the form of cognitive processes requires process models, but to date it is unclear which properties render a model a process model

— Process models require process data and behavioral data, and the avoidance of reverse inference from behavior to process

— Discovery of the function of cognitive processes requires an ecological rationality analysis (Gigerenzer, 1991), with content-rich norms (Arkes, Gigerenzer, & Hertwig, 2016), derived from evolutionary inclusive fitness of the cognitive system (Bischof, 1987)

— This opens up new avenues of research: the robustness of choice mechanisms, and the functional specification of preferential choice

— Regarding robustness: Human classification behavior under uncertainty about the probabilistic structure of the environment adheres to a simple yet robust principle called class-conditional independence

— Regarding functional specification: Human choices about uncertain outcomes differ in reproducible ways for different functions and utilize lexicographic cue integration from memory

I will specify the conceptual requirements for cognitive process models (Section 6.1), and then apply the form-function-centered analysis in two empirical studies: one on the problem of robustness in classification (Section 6.2) and one on functional domain-specificity in risky choice (Section 6.3).

1.1 The division: form or function in the behavioral sciences

The work on modeling the form and the function of decision processes is undertaken in different disciplines. Cognitive science investigates the form of contemporary decision processes (with few exceptions, see below). In contrast, evolutionary psychology, jointly with behavioral biology and ecology and part of economics, works out the functions of human choices.

1.1.1 Decision science: Cognitive models of behavior and process

Decision science analyzes in detail how the mind attends to, acquires, stores and retrieves, transforms, or integrates information. Researchers use a variety of models, ranging from highly complex neuronal networks to rather simple one-step algorithms. Since the 1940's, modeling the hypothesized information integration processes of the mind steadily expanded in cognitive science, and today, there is model competition and model proliferation: As of 2014 the decision making literature has accumulated more than 170 models[1], four cognitive modeling textbooks have appeared since 2010, and estimated 80 % of cognition-related publications involve a form of a model (Busemeyer & Diederich, 2010).

Categorization research epitomizes the multitude and content of decision sciences' models. *Prototype models* chose the class whose 'average' member resembles the new stimulus (Reed, 1972); *exemplar models* pick the class whose individual members are most similar to the new stimulus (Medin & Schaffer, 1978); *RULEX* combines if-feature-x-then-class-y rules into a decision tree (Nosofsky, Palmeri, & McKinley, 1994); *fast-and-frugal trees* are decision trees constrained to one exit option per node (Martignon, Katsikopoulos, & Woike, 2008); *ALCOVE* is a neural network (Kruschke, 1992), as well as *COVIS* (Ashby, Alfonso-Reese, Turken, U, & Waldron, 1998); then there are *rational models*, which utilize the conditional probability of classes given stimuli obtained through Bayesian inference (N. D. Goodman, Tenenbaum, Feldman, & Griffiths, 2008; J. R. Anderson & Matessa, 1990).

Two facts are noteworthy: First, no single modeling framework has emerged as the superior tool for predicting how the mind arrives at categorization decisions. Second, the questions regarding the robustness of categorization decisions across different structures of the environment is seldom asked (but see Luan, Schooler, & Gigerenzer, 2011, on fast-and-frugal trees).

1.1.2 Evolutionary psychology: Optimization of the function of modules

Evolutionary psychology explains choice mechanisms and behavior as evolved responses to recurrent evolutionary problems. To this end algebraic (e.g., Frankenhuis, Panchanathan, & Belsky, 2015), or agent-based (e.g., D. D. Johnson & Fowler, 2013) optimization models are employed to show that one strategy outperforms another in terms of expected reproductive value.

For example, Hintze, Olson, Adami, and Hertwig (2015) investigated whether risk-aversion as genetically inherited trait can emerge in groups of different sizes. Their agent-based simulation found that particularly in small societies, individuals evolve to avoid options with low-probability high

[1] According to a review of the decision making literature from 2004 to 2014; for details see Chapter 7.5.

outcomes but prefer less variable outcomes with a lower expected payoff compared to the more risky option (risk aversion). This is because on a population level, the society is more likely to die out before the improbable high-value outcome of the risky option will be realized for the first time. By contrast, selecting the high-probability but low-value option provides smaller societies with more reliable supplies. Only in large societies, a risk-favoring strategy (i.e., more agents in the population select the risky option) emerged.

The agents in such simulations are overly simplistic, because the research question concerns the stability of the final decision making *behavior*. It does not model the *process* producing the behavior. This would require using plausible cognitive process models as agents in such simulations. I am not aware that this step has been taken yet, but some work has been done to use plausible processing approaches for evolutionary functional mate choices (G. F. Miller & Todd, 1998). Yet, calls for integrating cognitive process models and behavioral biology decision rules (Hutchinson & Gigerenzer, 2005) have remained largely unaddressed, despite the conclusion by McNamara and Houston (1992) that "an understanding of the evolution of mechanisms requires a fundamental change in the sort of models that are analyzed" (p. 673).

1.1.3 Ecology and biology: Optimization models of evolutionary stability

Human behavioral ecology is the formal study of decision rules in different anthropological contexts. Their models are typically reductionist mathematical optimization methods. A model of "mechanism" refers not to cognitive processing but to evolutionary selection, like kin selection or sexual selection. Behavioral ecology formalizes the adaptive problem and costs "as simple as possible" (Winterhalder & E. A. Smith, 2000, p. 52). The same simple-rules optimization approach is used in behavioral biology (McNamara, Houston, & Collins, 2001), for example to model animal foraging (McNamara & Houston, 1992).

One core question in this field is: How does the organism maintain a stable structure against perturbations of their ecological niches (Kitano, 2002; Krakauer, 2006)? Arguments about the robustness of cognitive processing are largely absent in the form-focused studies, except for Gigerenzer, Todd, and the ABC Research Group (1999), in cognitive science, but in computer science and machine learning there are arguments about the robustness of computing algorithms (e.g., for the robustness of class selection of classification algorithms, Domingos & Pazzani, 1997).

1.1.4 Behavioral economics: Paramorphic models of the influence of incentives

Behavioral economics is mainly interested in how incentive systems, like legal or market rules, influence aggregate behavior. Their models are 'paramorphic' (P. J. Hoffman, 1960). They abstract from the individual cognitive processes, and model behavior given different environmental payoff structures. One example are models of risk taking, like prospect theory (e.g., Kahneman & Tversky, 1979; Tversky & Kahneman, 1992). The goal of behavior is maximizing monetary rewards or utilities, which economists model by optimization under constraints (Berg & Gigerenzer, 2010).

What behavioral economics and the biological approaches share is the focus on optimality or stability of choice strategies and the corresponding models; where they differ is the contents of the goals. The former define reproductive success as the goal, while the latter define monetary value or utility maximization as the goal.

1.1.5 Summary

Models of form and function of human choices differ with respect to the detailed specification of cognitive processes and whether robustness of a strategy is of interest. Specifically, research on the function of choices or the underlying processes largely uses reductionist, simple models, whereas the work from cognitive psychology aims to specify the mental processes in greater detail.

1.2 An integration: Form follows function

In a review of the fundamental problems for decision making research, Hastie (2001) called for investigation of whether the form of cognitive processing is shaped by evolutionary functioning given different problem domains.

In my dissertation I take the view of integrating form and function, a view which has been repeatedly emphasized historically. Brunswik's (1955, 1956) lens model is a theoretical framework according to which decision makers use proximal cues (i.e., the information the system processes) that correlates highly with distal variables (i.e., the to-be-achieved goals). Gigerenzer and colleagues proposed the concept of ecological rationality (Gigerenzer, Todd, & the ABC Research Group, 1999; Gigerenzer, Hertwig, & Pachur, 2011): behavior is ecologically rational if it results from decision rules (heuristics) that perform well with respect to current goals by relying on the information structure in the environment (i.e., the mind applies the decision strategy appropriate for the context). Bischof's (1998) systemic analysis prescribes that the structures of the cognitive system is adapted to use information in order to perform evolutionarily relevant functions. Finally, Anderson's (1990) rational analysis concerns the optimality of behavior given environments and goals.[2] These various approaches share that they aim to integrate form and function.

Despite their similarities, these approaches are distinct. The first distinction is that the outlined proposals differ with respect to what constitutes goals. J. R. Anderson (1990) focusses on proximate, contemporary goals. Gigerenzer, Todd, and the ABC Research Group (1999) allow for many different goals but state that the core cognitive capacities underlying the decision processes have evolved; whereas Bischof (1998) proposes to study ultimate goals. Proximate goals are part of the current environment, like accurate medical diagnosis; ultimate goals are indirect, like maximizing inclusive fitness (reproduction of the own genes). I advocate starting with the study of evolutionarily relevant goals.

[2] A terminological note: Brunswik and the evolutionary literature use the terms *proximal* and *distal* differently. For Brunswik distal variables are goals and proximal variables is the direct stimulation of the organism (e.g., distal goal: to perceive the color red; proximal variable: light-waves on the retina) (e.g., Brunswik, 1943). For evolutionary psychologists distal variables are fitness-relevant goals and proximal variables are contemporary goals of an organism (e.g., Haselton, Bryant, et al., 2009).

Why evolutionary goals? Evolutionary functions provide behavioral norms without arbitrary coherence to logics, mathematics, or statistical principles. Arkes et al. (2016) argued that performance trumps coherence if coherence concerns "content-free norms" (p. 34). One way to instill content into norms is to link them hierarchically to survival or reproduction (for an information theoretic argument, see Bischof, 1987). Even though it is clear that humans can perform well in tasks without evolutionary functioning — they solve mathematical puzzles, judge city sizes, and play jeopardy — the question is whether it is a useful starting point to study the mind from such non-evolutionary goals. Not every behavior that results in some performance needs to have evolved and not every evolved behavior performs a task well. It is uncontroversial that the basic mental operations recruited to play jeopardy have likely evolved to cope with evolutionary problems. However, uncovering cognitive processes requires some degree of conditional stability of the process. That is, if every person has an idiosyncratic strategy for jeopardy the resulting data will be noisy. Thus, it seems a good starting point to select a task for which the expected noise in strategies is low. Evolution thus provides a starting point to define what processing steps there are, and how they could work given evolutionarily relevant input conditions. This answers the question: Do some people employ a certain strategy at all? In the next step, one can ask in which current tasks the discovered processing steps are recruited. This answers the question: Do people use the available strategies in an ecologically rational way?

Another distinction between the proposals for form-function integration outlined above is that they differ with respect to which type of model adequately formalizes the cognitive system. J. R. Anderson (1990) proposes Bayesian models; Bischof (1998) proposes system's theoretic models; Gigerenzer, Todd, and the ABC Research Group (1999) proposes simple heuristics. I advocate to use a class of models — process models — which encompasses some of the aforementioned model types.

Why process models? The reason for process modeling is that there are two goals: One is to find out which *form* cognitive processes take, and another one is to simultaneously constrain the variability of the possible forms by its *functions*. Therefore, we need process models that can be tested on both a process and outcome level. One of the problems is that it seems unclear which properties render a model a process model.

Next I will outline implications of the form-function integration.

1.2.1 Process models

The research questions about optimality and stability of choice rules need to be addressed by analyzing the space of process models with respect to the goal, instead of optimization over reductionist models. To date, the agents in agent-based models are usually engineered by 'forward engineering' which postulates making the agent as simple as possible. Usually this entails one equation per agent. Making agents slightly more complex would lead to, for example, simulations of the evolutionary constraints under which a fast-and-frugal decision tree outperforms an exemplar process in classifying nutrition density of food. This requires introducing process models into the evolutionary literature. One obstacle to the introduction of process models is the current model proliferation in cognitive science, which I outlined for categorization research above. Another obstacle toward process modeling consists

in that, especially for researchers unfamiliar with the cognitive literature, it can be hard to disentangle which models constitute process models.

1.2.2 Robustness of learning

The second implication of a form-function integration is that learning mechanisms are constrained to perform functions robustly. Learning presumes an evolved cognitive structure that utilizes signals and goals in the environment to make decisions. System's biology and ecological rationality both stress that evolved structures need to perform their function robustly to withstand perturbations by the environment (Krakauer, 2006). Cognitive systems therefore are selected to function across many circumstances. Selection means that variance in fitness reduces the variance of cognitive processes. This is because the processes that produce behavior which results in low fitness will not reproduce: the function constrains the form. The cognitive processes for learning are themselves evolutionary selected. Food preferences are a case in point. Rats learn to avoid food that was followed by nausea much quicker than they learn to avoid sounds followed by pain (Garcia, Hankins, & Rusiniak, 1974). Therefore we expect that humans, by default, use rather robust cognitive strategies — at least in the absence of other experience.

1.2.3 Functional specification of information integration

If functionally relevant goals inform the form of decision making, one is required to study decision making in multiple contexts with different functions. According to ecological rationality (Todd et al., 2012; Arkes et al., 2016), integrating form and functions involves asking which strategies achieve goals given which environmental structures. The evolutionary approach predicts that the recruited cognitive processes should vary least for goals with high fitness validity, and be more heterogeneous as the fitness validity of the goal decreases. For example, Fischbacher, Hertwig, and Bruhin (2013) find different classes of behaviors in economic experiments about allocating money. Abstract, numerical, purely monetary stimuli have low fitness validity if we consider that in the ancient environment trade was personal, goal-directed, and driven by goods rather than an intermediary commodity. Mones is regarded as "not directly adaptive" (Lea & Webley, 2006, p. 162). The evolutionary integration of form and function presumes more homogeneous cognitive processes if the process is tested on goals with high fitness validity (note that different processes with equal performances are permitted). What follows is a functional specification of cognitive processes across fitness-relevant goals.

1.2.4 Preferences as inferences

Combining proximate and ultimate goals is key for merging form and function. The dual-functional study of cognition implies that preferences for goals can be studied as inferences for fitness validity. Preferences for attractive partners or partners with social status become inferences for reproductive value (G. F. Miller & Todd, 1998), while preferences for sweet and salty snacks become inferences for nutrition density (Birch, 1999). The underlying processes can be studied to achieve these functions.

That is not to say that every preference is evolved or everything evolved into preferences. Ontogenetic learning modulates cognition. The evolutionary functional perspective, however, provides a starting point for modeling preferential choice processes.

Chapter 2

Projects and theoretical background

This chapter outlines the background for one theoretical and two empirical projects to further form-function integration in decision science.

2.1 Models of the form: Process models

What the aforementioned meta-theories of form-function integration (Brunswik, 1955; Gigerenzer, Todd, & the ABC Research Group, 1999; Bischof, 1998) share is connecting a mathematical description of the cognitive process with a measure of achieving performance goals. There are multiple views on how to formalize the cognitive process: while Gigerenzer, Todd, and the ABC Research Group (1999) advocated simple heuristics, Bischof (1998) favors system-theoretic models. However, even though specific model classes are well-specified (e.g. the search-, stopping-, and decision-rule of a heuristic, Gigerenzer & Sturm, 2011), the generic class of 'process model' is ambiguous. In the introduction to their cognitive modeling textbook, Sun (2008) encouraged the exploration of the "design space" of models (p. 15). Thus: What is a cognitive process model?

In decision science process models are prominent. Between 2005 and 2015, the term "process model" has appeared in the text of roughly 12,400 scientific documents from cognitive psychology. The citations of database-indexed papers using the term have increased steeply (even when controlling for a positive citation trend; own analysis) and there has been a corresponding growth in interest in process tracing measures (Schulte-Mecklenbeck, Kühberger, & Ranyard, 2011, p. 9).

However, textbooks offer limited instructions for process model development or the necessary characteristics required for process models. The textbook by Lewandowsky and Farrell (2010) advises that parameters in a process model need a psychological interpretation (p. 18), and that process models need to describe the process in detail (p. 25). A review of the modeling literature found little overarching guidance regarding which aspects matter for process models (see Chapter 4 below).

Therefore, I surveyed experts and conducted a literature review about the properties of process models. The elaborations above suggest a lack of clarity regarding process model characteristics.

2.2 Robustness of learning: A simple but robust categorization system

The first integration of form and function of choice concerns categorization. Categorization involves assigning objects according to their features to classes. According to the behavioral ecology and

ecological rationality literature, a cognitive system, if it is functional, needs to be robust against environmental changes (Kitano, 2002; Gigerenzer, Todd, & the ABC Research Group, 1999), which is particularly relevant considering the computational complexities of categorization.

Categorization is among the most-studied cognitive abilities, as the brief model review in the introduction showed. Humans can learn many arbitrary category structures provided enough feedback (e.g., Medin & E. E. Smith, 1981; Little, Nosofsky, Donkin, & Denton, 2013; Nelson, 2005; McDaniel, Cahill, Robbins, & Wiener, 2014); they can even learn to classify galaxies (Lintott et al., 2008). The system *performs*, yet single-task performance cannot tell us about the robustness of the categorization system across different task structures. How robust is human categorization?

Classification implies at least two computational challenges. One is the *curse of dimensionality* (Bellmann, 1961), which refers to the fact that the number of feature-category combinations grows exponentially in features, e.g. for binary features the number of combinations equals $2^{N(\text{features})}$. This is relevant especially in the real world with more than a few features. Another challenge is *inferences from little data*. It means that, although a person may not have experienced every feature combination, people can extrapolate from training to transfer exemplars (e.g., McDaniel et al., 2014). Robustness means that the categorization system is accurate given many different task structures.

Is there a simple but yet robust solution? A probabilistic classifier called Naïve Bayes was studied in machine learning for both these properties. It treats features as statistically independent given the true class, which is known as the *class-conditional independence assumption*. Making this assumption can lead to accurate classification decisions even if in fact features are not independent given the classes (Domingos & Pazzani, 1997; Rish, Hellerstein, & Thathachar, 2001). Class-conditional independence reduces the curse of dimensionality by reducing the number of required parameters a probabilistic classifier needs to learn from an exponential growth to a linear growth. Further, it enables inferring the class for previously unseen feature combinations if the individual features were observed in a different combination in the past. Yet, the class-conditional independence assumption has limits. For example, it cannot learn exclusive-or problems in which the objects that differ in all features belong to the same class. Humans, however, can learn exclusive-or classifications (e.g., Little & Lewandowsky, 2009).

I hypothesize that the classification system adheres to this robust principle, but not rigidly. People use class-conditional independence as default assumption at the beginning of learning, but can adapt their classifications to different environmental structures. I designed two experiments and one model of classification learning to test this hypothesis explicitly.

2.3 Functional specification: Survival as function of risk information processing

The next study focuses on functional risk taking behavior. Risk taking involves choosing an option with variability in outcomes, like: Will I eat the marinated cockroaches at the new experimental restaurant? According to an evolutionary view on risk information processing, the mechanisms underlying risk taking should be adapted to different functional goals. This study asks: Which cognitive processes are

responsible for domain differences in risk taking?

Note, that the notion of risk in the evolutionary literature corresponds to Knight's notion of uncertainty, where probabilities may not be precisely known (Knight, 1921).

There are two main views concerning the generality or specificity of human risk-taking behavior. First, such choices may result from relatively general rules, which is a position is held by many economists (Kahneman & Tversky, 1979; Einav, Finkelstein, & Cullen, 2010) and cognitive psychologists (J. R. Anderson, 1990; Busemeyer & Bruza, 2012; N. H. Anderson, 2014). Alternatively, risk taking could be the result of choice processes adapted to the goals the cognitive system aims to meet with taking risks. This position is held by some cognitive psychologists (Gigerenzer, Todd, & the ABC Research Group, 1999; Weber, Blais, & Betz, 2002) and many evolutionary psychologists (e.g., Cosmides & Tooby, 1992; X.-T. Wang, 1996; Barrett & Kurzban, 2006).

The work on domain-specificity faces two limitations. The first is an ad-hoc choice of domains (Kühberger, 1998). The number of domains varies from four to ten (Rettinger & Hastie, 2001; Weber, Blais, & Betz, 2002; Wilke et al., 2014), and so do their labels. Optimally, the content of domains should be identified from a theory that is largely external to the cognitive literature. Wilke et al. (2014) recently used evolutionary theory to identify ten domains based on a literature review on the functions of evolutionary risk taking (for details, see section 4). This is one possibility to address the domain-selection problem.

The second limitation stems from methodological narrowness. Domain-difference studies often focus on psychometry (Blais & Weber, 2006; Wilke et al., 2014). While scales are necessary to establish domain-differences, they are limited regarding how people arrive at their choices. What follows is a lack of process tracing across domains (most process-tracing studies concern single domains, e.g., Montgomery, 1977; Cokely & Kelley, 2009; Payne & Braunstein, 1978; one exception is Rettinger & Hastie, 2001). One reason for the lack of process tracing is the format of questionnaires; another is that it is unclear what the to-be-processed cues in particular domains are.

My study traces the cues people use across functionally specified domains. I was interested in establishing whether there are domain differences in risk propensities, and whether selecting domains evolutionarily results in stable differences in risk taking across domains. I was also interested in whether the direction of the risk-cues that people retrieved (pointing towards more or less risk taking) or the content of the cue relates to differences in peoples' risk propensities across domains.

Specifically, I asked whether the retrieved cues are combined in one of three ways. First, people may combine cues by a simple *count of positive vs. negative cues*. Tally (Dawes & Corrigan, 1974) is an information processing heuristic (see also Gigerenzer & Goldstein, 1996) that counts the number of cues in favor of an option and picks the option with most positive cues. Second, people may combine cues by a non-compensatory process where the cues that are retrieved earlier dominate later cues; thus, the *order of retrieval* of positive and negative cues would determine risk propensities. Non-compensatory processes are implemented in a number of heuristic models (Goldstein & Gigerenzer, 1999; Tversky, 1972; Brandstätter, Gigerenzer, & Hertwig, 2006) and theories (E. J. Johnson, Häubl, & Keinan, 2007). Third, people may retrieve *qualitatively different cues*, independent of the direction

selectively in different domains. Previous work showed that people use particular subsets of all available information that are valid predictors for a good choice, even if more information is available (implemented in the take-the-best model, Gigerenzer & Goldstein, 1999).

CHAPTER 3

HYPOTHESES

This section presents the predictions for process models, the robustness of learning, and functional specification.

3.1 Process model: A concept with many meanings

Regarding the use of process models, I hypothesize that there is no consensus in cognitive science regarding process model properties:

H1a Experts judgments disagree about the properties of process models

H1b Definitions in the literature disagree about the properties of process models

3.2 Robustness: A functional property of categorization behavior

Regarding the robustness of categorization decisions, I hypothesize:

H2 People's initial classification decisions follow the robust principle of class-conditional independence. Classification learning is best described by model (DISC-LM) with a high prior belief in class-conditional independence

3.3 Survival: A functional property of risky choice

Regarding the functional specification of choices about risks, my first hypothesis concerns the dependency of risk taking propensities on evolutionary functions and the replicability of these differences.

The next hypothesis concerns alternative accounts of cue integration underlying the domain differences in risk taking. Let positive cues denote risk-favoring ("If x, I would be *more* likely take risk y") and negative cues denote risk-avoiding cues ("If not x I would be *less* likely to take risk y").

In sum, I hypothesize:

H3a The propensities to take risks differ across domains of various evolutionary functions, replicating previous findings

H3b In domains with higher risk propensities, more positive cues than negative cues are retrieved

H3c In domains with higher risk propensities, positive cues are retrieved before negative cues

H3d In domains with higher risk propensities, the specific cues are retrieved, independent of the cue direction

Chapter 4

Materials, procedure, and methods

The next section describes the employed materials.

4.1 Process Models: Online survey, literature review, and reliability analysis

To measure scientists' opinions regarding process models I used an online survey asking participants to classify 116 models (identified in a literature review, see below) as process models (yes — no — no opinion). The survey also asked participants if the algorithm in Marr's (1982) three levels of analysis clarifies the process models, and for a definition of process models, followed by demographics and level of experience (professor — researcher — student).

Agreement about process model properties was treated as if the scientists were raters in a qualitative data analysis. Fleiss-Cuzick's $\varkappa$ was used as inter-rater statistic. The reliability measure is suitable for the present data: dichotomous ratings by more than two judges with an unequal number of judges per item (Fleiss & Cuzick, 1979).

To determine how the literature conceptualizes process models, I identified decision making models in the literature. Models were selected based on novelty and importance, content, and usage. A broad search for important or novel papers using cognitive modeling (see sampling methods below) identified a preliminary list of articles from which I obtained all tested models. From the models, I selected those with contents related to decisions. I categorized whether the models were about decision making or other topics (e.g., attention), and we performed inter-coder analysis (between Jolene Tan and myself) on a random subset of 50 models (Cohen's $\varkappa = .831$, we discussed disagreements and I adapted my initial categorization). I further excluded models of low relevance in the field: I looked at whether the articles in which the models were originally proposed were still cited, and included only models with a source that had > 388 citations in total, or > 6.6 citations per year.[1] The resulting list of models whose properties were reviewed had 116 entries.

[1] 388 is the 66th percentile cutoff of all citations, and 6.6 is the 33rd percentile cutoff of average citations per year in Google Scholar.

4.2 Robustness of categorization: Laboratory experiments, optimal experimental design, and individual cognitive modeling

To investigate categorization learning I specifically designed the task in order to achieve high experimental control. For the same reason, the studies were conducted in the Max Planck Institute laboratories. I used a computerized trial-by-trial supervised category learning study.

Participants' task was to learn to classify pictures of a plankton specimen into species A or B. In each trial they saw a randomly drawn exemplar, classified it, and received feedback. Learning ended after a performance criterion was reached in order to capture individual differences in learning speed. The experiment used a biological cover story because it is rather natural, but people have little experience with plankton.

I designed the structure of the underlying classification task to maximally discriminate our hypothesis (H2: People's initial classification decisions follow the robust principle of class-conditional independence. Classification learning is best described by model (DISC-LM) with a high prior belief in class-conditional independence), using optimal experimental design principles (I. J. Myung, Balasubramanian, & Pitt, 2000; Nelson, 2005). This involves seeking parameters — the probabilities of the class and the objects within classes — such that the formally predicted behavior differs maximally with respect to the hypothesis. These designs can discriminate models better than balanced, D-optimal, designs (I. J. Myung et al., 2000); but are less realistic than representative design (Brunswik, 1943).

Numerical optimization methods with a genetic algorithm were used obtain task parameters that differentiated our hypothesis (because the underlying optimization problem has no analytical solution). The resulting task structure violated class-conditional independence strongly. Further, exemplars were deterministically associated to classes, which matters in learning experiments because deterministic categorizations tend to be easier to learn than probabilistic ones (Little & Lewandowsky, 2009; Mehta & Williams, 2002; but see Juslin, Olsson, & Olsson, 2003). I therefore manually changed some probabilities to obtain a design with an identical but non-deterministic class membership, leaving me with two statistical environments.

To analyze whether people presume class-conditional independence in category learning, I developed the dependence-independence structure and classification learning model, DISC-LM. It formalizes a learner who is uncertain about whether features are statistically independent given the class. Intuitively, the DISC-LM learns and utilizes two aspects of the world, *probabilities* and *structure*. Probability-wise it learns the probabilities of classes and of exemplars in classes, which it needs to compute the classification decision in each trial. Structure-wise it learns whether the individual features that make up exemplars are statistically independent given the true class. Learning was formalized as Bayesian updating.

Importantly, the model formalizes a *prior belief* about whether the task structure follows class-conditional independence. If the value of the prior belief in class-conditional independence changes, the model changes its classification behavior. As the model experiences the task it updates this prior belief and learns whether class-conditional independence actually holds.

To see how model behavior depends on the prior belief in class-conditional independence I simulated the

model with various values of the prior belief. I employed Monte Carlo methods to obtain probability densities. The model was presented with the first 50 trials of the exact same sequences that our participants had seen in the experiments. Then I looked at the learning curves, in the aggregate, and compared whether the pattern matched the learning curves of the human subjects.

To investigate whether individual participants start classification learning with a belief in class-conditional independence, I used one-trial-ahead predictions of individual data (predicting each individual's next decision given the previous experience) and used mean squared error (MSE). MSE emerged as the most reliable measure in a parameter recovery study (compared to absolute errors and likelihood-based measures). I was interested in whether a high parameter value for the prior belief in class-conditional independence described individual choices best.

4.3 Functions of risk taking: Online survey, process tracing, statistical modeling, replication analysis, and qualitative content analysis

To address domain differences in risk taking, I conducted an online survey.

Risk propensities were measured with the evolutionary risk scale developed by Wilke et al. (2014). The scale items were derived as modern-day behaviors corresponding to fitness-related goals which in turn were defined from a review of the anthropological and evolutionary psychology literature. The 30-item scale assesses how often people engage in risky actions on a seven-point Likert scale (extremely unlikely — extremely likely). There are three questions for each of the ten domains: Within-group competition, between-group competition, environmental exploration, status/power, kinship, parental investment, food acquisition, food selection, mate attraction, mate retention. Risky behavior regarding food-acquisition involves, for instance, *eating a piece of food that has fallen on the floor*.

Further, the following life-history variables were measured: Age, gender, relationship status, number of siblings, birth order, number of offspring, minimum and maximum number of desired offspring, and life expectancy.

To obtain the cues that people recalled a proces tacing technique known as aspect listing was employed (E. J. Johnson, Häubl, & Keinan, 2007). It involves asking people to list all relevant aspects on their mind related to a choice and also the direction towards which this aspect changes choices. The method surveys the aspects one-by-one, thereby also measuring the retrieval order. The survey asked people after they had responded to a risk scale item to report situational aspects that would increase or decrease their risk taking behavior in the given situations. People could list up to ten such cues.

I analyzed the data from the risk survey with ordinal logit regressions. Because respondents treat inner and outer scale points on Likert-type scales not as equally distant from each other (e.g., Hamby & Levine, 2015; Lodge, 1981; Cronbach, 1946; Lantz, 2013), the data from the risk survey have an ordinal measurement level. Analysis of variance (ANOVA) has severe limits if applied to ordinal data (remarked for the binomial case already by Cochran, 1940; for summaries see Agresti, 2002; T. F. Jaeger,

2008). I employed ordinal multinomial logistic regressions[2] to analyze the effect of domains and gender on risk taking likelihood. I included gender for it is one of the most-discussed variables in the risk literature (Byrnes, Miller, & Shafer, 1999). Further, risk taking likelihood was a repeated measure (each person reported it for ten domains), thus I used a mixed models to take the within-person response correlations into account. All models specified participants as random effects and other variables as fixed effects. I used the Akaike information criterion (AIC Akaike, 1974) to assess whether including domain increased model fit, comparing a full model ($likelihood \sim domain \times female$, where $\sim$ denotes "is regressed on") to a restricted gender-only model ($likelihood \sim female$) .

I conducted a replication analysis by comparing how many of the obtained effects in our non-student sample pointed in the same direction as the effects from running the same model on two studies with students that used the same risk scale, study 2 and study 3 by Wilke et al. (2014).

To analyze the statements about which cues would increase risk propensities, I followed qualitative content analysis principles (Mayring, 2014) (using www.qcamap.org). Our coding units were individual statements, the analysis goal was individuating the types of cues, the analysis was directed at the written content (rather than its effect on the reader), and it had two sub-goals: to classify the types of cues, and to classify the direction of the cue, i.e. whether it was a positive (risk-favoring) or negative (risk-avoiding) cue. I generated ten coding manuals (one per domain) from part of the data. A research assistant cross-coded. She was trained in three training sessions on 20 to 26% of responses per domain, I amended the coding manual if necessary, then both raters coded the remaining data independently. I computed inter-rater reliabilities, and then preprocessed the data, i.e. resolved coder discrepancies by discussion, and excluded erroneous statements.

The inter-rater reliability statistic were suitable for two raters, multiple nominal categories, and unequal marginal category distributions. Our key measure is Gwet's AC_1 (Gwet, 2008), which performs particularly well when categories are unequally frequent (which was the case in the present data). As a robustness check I also computed other measures: Krippendorf's α (which can be outperformed by Gwet's AC for extremely uneven marginal category distributions; Gwet, 2008), and Brennan-Prediger's $\varkappa_n$ (which does not correct for unequal marginal categories; Feng, 2014); as well as Cohen's $\varkappa$ (because it is the most-widely used measure).

[2]Intuitively, this method computes the odds of responses less or equal than the j^{th} point of the likert scale compared to responses above j for each domain ($j = 1, \ldots, 6$). It then uses the ratio of two odds from two domains as a measure of whether risk taking likelihoods differs between domains (for details see Agresti, 1989).

CHAPTER 5

SAMPLING AND DATA SOURCES

All empirical investigations were conducted in accordance with the ethical and data protection guidelines at the Max Planck Institute for Human Development, Berlin.

5.1 Process Models: Databases for the review, decision science sample

For the literature review on process models, I searched the databases Google Scholar and ISI Web of Scienc for publications (a) appearing between 2004 and 2014, OR with a citation number greater than 100, (b) published within *Psychological Review, The Journal of Experimental Psychology: General, The Journal of Experimental Psychology: Learning, Memory, and Cognition*, or *Judgment and Decision Making* which (c) included the term "model of" AND synonyms[1] for the term decision making.

For the survey about experts' opinions on process models, I contacted scientists by, on the one hand, directly emailing the developers of the models identified in our review, and, on the other hand, sending a call to the field's biggest mailig list (www.sjdm.org) and a specialized process tracing mailing list (www.egproc.org).

Our sample consisted of 62 people, 35 professors, 16 post-doctoral researchers, and 11 doctoral students. The survey asked respondents, among others, for their personal classification of the 116 models identified in our review as process model, no process model (yes/no/no opinion). The obtained classifications were analyzed in terms of inter-rater agreement.

5.2 Robustness of categorization: Laboratory participant sample

For the study on categorization learning, I sampled participants from the Max Planck Institute's participant pool.

In the first experiment 30 people participated and none were excluded. Participants were between 19 and 33 years with 20 female (mean age 23.8, $SD \pm 3$, range 19 to 33 years, 67 % female). In the second experiment 39 people participated and ten were excluded (eight did not reach the learning criterion within ca. 120 minutes, two due to a computer crash during the session), leaving us with a sample of 29. Participants were 18 to 35 years, and 23 female (mean age 24.8, $SD \pm 4$, range 18 to 35 years, 79 % female). They received 12 Euro as compensation in both experiments.

[1]The precise search phrase reads "model of * decision" OR "model of * decisions" OR "model of * choice" OR "model of * choices" OR "model of * preference" OR "model of * preferences" OR "model of * inference" OR "model of * inferences" where OR denotes the Boolean OR and * can be any word.

5.3 Robustness of categorization: Amazon Mechanical Turk sample

For the study on risk attitudes, I aimed to sample a diverse, non-student, population from North America, because the scale used was developed there. Participants were recruited through Amazon Machanical Turk's (AMT) online crowd-sourcing service, with localization restricted to North America. Participants from AMT behaved similar to laboratory participants in many cognitive-behavioral tasks, except for learning studies or visual priming (Crump, McDonnell, & Gureckis, 2013; J. K. Goodman, Cryder, & Cheema, 2013). As recommended the study included attention check questions.

One hundred and twenty six people participated, six were excluded (due to inconsistent responses), leaving a final sample of 120 (mean age 33.4 years, $SD \pm 11$, range 18 to 65 years, 52% female); they received 2 US dollars for participation (the study lasted on average 22 min, range 7 to 57 min).

CHAPTER 6

RESULTS AND DISCUSSION

6.1 A Framework for Cognitive Process Models (Jarecki, Tan, & Jenny)

This chapter demonstrates that the tool "process model" is ambiguously defined, by scholars and in publications alike, and proposes a framework specifying the properties of (general) cognitive process models.

6.1.1 Ambiguity of process model properties

The first questions concerns whether process model properties are clearly specified — either implicitly by expert opinions, or explicitly in the literature.

H1a: Experts judgments disagree about the properties of process models. As expected, experts who classified the decision science models judged different models as process models, with a low inter-rater agreement (Fleiss-Cuzick's $\varkappa = .27$ with values above .60 regarded as 'good' agreement; Fleiss & Cuzick, 1979). Classifications by professors were slightly more consistent than researchers or students, but still low ($\varkappa = .33, .17$, and .14, respectively). This suggests that meta-theories related to process models, most prominently Marr's (1982) three levels of analysis (algorithm, computation, implementation) have underspecified the properties of process models (despite their frequent use, e.g. Chater, 2009; Griffiths, Lieder, & Goodman, 2014; K. Huang, Sen, & Szidarovszky, 2012; Jones & Love, 2011; McClelland et al., 2010). When asked if the algorithmic level clarifies what process models are, the 38 people familiar with Marr were divided equally between the extremes "does not clarify at all" and "claraifies completely" ($\chi^2(2) = 14.105$, $p < .001$, Cramer's $V = 0.431$). This means that there is a lack of clarity about process model requirements among researchers.

H1b Definitions in the literature disagree about the properties of process models. The review uncovered that process model as a method appears in different contexts. In part of the literature process-type models are tested and compared against *rational-type models*. These models describe choices that either solve certain tasks optimally (rational models, J. R. Anderson, 1991a), or that reach optimality within a fixed margin of error (rational process models, Sanborn, Griffiths, & Navarro, 2010). In comparison, process models are seen as less flexible and rather mechanistic tools.

The second context sets process models in contrast to *as-if-type models*. These models assume either unrealistic mental operations (as-if models, cf. Berg & Gigerenzer, 2010), or leave open whether the computations are realizable as long as the functional relationships capture behavior well (paramorphic models, P. J. Hoffman, 1960). In comparison to these models, process models are seen as more realistic

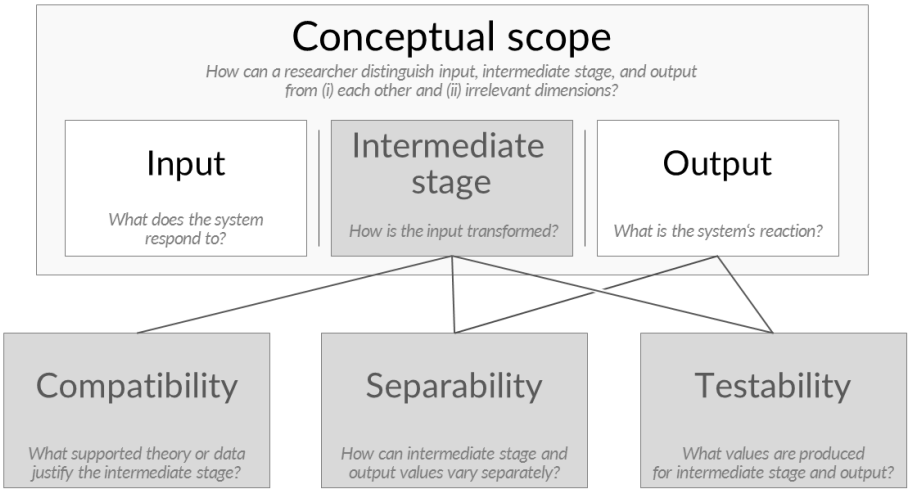

Figure 6.1. A framework for cognitive process models. Grey boxes denote properties specific to process models; lines denote the interdependencies between the features of process models.

and data-constrained tools implementing, for example, cognitive capacity limits. Note, that also rational models make no assumptions about underlying processes, thus, the first two connotations overlap.

Thirdly, 'process model' refers to a class of *dynamic, stochastic models* of cognition (e.g., diffusion, or accumulator models, Busemeyer & Townsend, 1993; Usher & McClelland, 2001). These tools typically share one formal feature: random walk processes. In this context, process models share a formal feature. These different contexts may amplify the lack of clarity regarding what constitutes a process model.

This suggests that the literature features a variety of characteristics required for process models.

6.1.2 A Framework for process models

To address the lack of clarity, I propose a framework for cognitive process models: As shown in Figure 6.1 a process model includes (a) one or more intermediate processing stages, (b) is built such that the intermediate stages are compatible with current knowledge about human cognition, (c) constructs the intermediate stages such that they can vary separately from the output, and (d) allows testable claims to be derived about both the intermediate stages and the output, given the same input. In addition, a tri-modal conceptual scope is necessary.

Tri-modal conceptual scope. The scope describes the phenomena to which a model applies qualitatively. The tri-modal scope that process models require include inputs, outputs, and to which mental operations it applies (e.g., attention, neuronal activation, or causal beliefs). Gluth, Rieskamp, and Buchel (2012) model the decisions to acquire goods. Their model takes the good's value as inputs, processes at the neuronal level, and outputs rejection/acceptance decisions. This scope is tri-modal. By contrast, Fehr and Schmidt (1999) model contributions in economic games based on fairness, but

fairness is defined in terms of the output.[1]. This is a bi-modal scope. Would fairness be defined as emotional or neuronal phenomena the scope would be tri-modal. The tri-modality requirement is deliberately broad. Of course, a conceptual scope may change in subsequent applications of a model — important is a reference to the cognitve processing phenomena.

At least one intermediate stage. In addition to a tri-modal scope, process models have at least one intermediate stage specified (in line with Svenson, 1979; Weber & E. J. Johnson, 2009). Intermediate stage are precise mental operations that—according to the model—transform input to output. A single intermediate stage may be attention distribution, neuronal activation pattern, or the structure of causal beliefs. Several intermediate stages may be sequences like a change in attention, neuronal firing, or network activation. A model without intermediate stages is cumulative prospect theory (Tversky & Kahneman, 1992). It formalizes the utility u of receiving x with probability p as $u(x,p) = v(x)w(p)$, where $v(x)$ and $w(p)$ are valuation and weighting functions. It under-specifies the order of operations (weighting before, after, or concurrent with valuation?). Kahneman and Tversky (1979) initially hypothesized weighting before valuation, which is missing in the model. Models containing equations can, however, have intermediate stages if the equation explicates the temporal order, like in sequential sampling or drift-diffusion models (Busemeyer & Townsend, 1993). Importantly, intermediate stages require a tri-modal scope.

Compatibility. The transformations in the intermediate stages of a process model should be, by and large, compatible with our knowledge of cognitive capacities. It requires either relating the intermediate stages to supported cognitive theories, or relating it to data about cognitive capabilities. For example, implementing an attention distribution that does not violate established bottom-up influences, or showing that a modeled belief network is not computationally intractable. Compatibility refines calls for plausibility of process models (e.g., Winkel, Keuken, van Maanen, Wagenmakers, & Forstmann, 2014), where the meaning of plausibility tended to be subjective (for a potential list of criteria, see Gigerenzer, Hoffrage, & Goldstein, 2008).

Separability. Process models yield at least two separate hypotheses given the same inputs, one hypothesis about outputs and, at least one about intermediate stages. For example, a process model of medical choices could jointly model the state of a causal network and the resulting diagnosis given the symptoms. The Take-the-best model (Gigerenzer & Goldstein, 1996) derives both the order of information acquisition and the resulting decision from cue values and validities.

Separability allows output and process to be disentangled, while they remain connected through the model and theory where outputs result from intermediate stages, and intermediate stages result from inputs. Processes cannot be caused by outputs (except in models with feedback loops). Separability prevents reverse inference (concluding from supported behavioral predictions that processes are true; i.e., affirming the consequent[2]), which is problematic due to the under-determination of outputs by processes.

[1] As the difference between the own payoff and the average of everyone else's payoff.

[2] Interpret P = process and B = behavior. The logical fallacy makes the following invalid inference: (1) Let P then B be true. (2) Let B be true. (3) Therefore P. The valid inference is: (2) Let $not B$ be true. (3) Therefore $not P$. (Geis & Zwicky, 2011, p. 562).

Testability. Testability means that the claims for outputs and intermediate stages are specific enough to be tested within the conceptual scope. While most formal cognitive models can be tested regarding their outputs (behavior), process models need to yield additional specific hypotheses for any intermediate stages encompassed by their scope. For example, the precise parameters of the state of a causal network as well as the numerical probability to decide for an option. The evidence strength here is asymmetric: If a process model fails behaviorally its computations are invalid—independent of the support of its process predictions—but if a model performs behaviorally its computations may or may not be valid—depending on the support of its process predictions.

6.1.3 Discussion and limitations

Merging functional and formal aspects of human decision making needs cognitive process models and optimality models connected. However, to date it is unclear which properties cognitive process models possess. I proposed a general framework specifying the requirements for process models.

Importantly, it does not value process models over outcome models or optimality models. The framework aims at defining the term. In discussion with colleagues and at conferences, "process models" tends to be used like "good model". However, which model is good depends on its purpose and explanatory power, as well as the research question.

Explicating the properties of process models goes beyond terminological issues; it facilitates teaching process modeling, guides model development, informs conceptual integration, and instructs supplementing behavioral models with a process level. It permits solving debates on the status of models, and aids model-based process tracing.

6.2 Robustness of categorization: Naive but robust computations in categorization learning (Jarecki, Meder, & Nelson)

Our task involved four *critical* feature combinations for which, by design, class-conditional independence selected a class different from the true class (*000*, *001*, *010*, *100*), and one *uncritical* exemplar for which the class decision was identical (given an arg max choice rule[3]).

H2: People's initial classification decisions follow the robust principle of class-conditional independence. Classification learning is best described by model (DISC-LM) with a high prior belief in class-conditional independence translates into slower learning of the critical compared to the uncritical exemplars. The following results hold for both experiments.

6.2.1 Classification errors

The first analysis investigated the aggregate error rates. If class-conditional independence is used, we expect more overall errors for the critical compared to the uncritical feature combinations. In line with

[3]Selecting class 1 if $p(class1 \mid exemplar) > .5$.

this, participants made on average the most errors for the critical feature combinations, and the least errors for the remaining exemplar. Though this supports H3b, the average error rates aggregate over individuals and trials, and collapse whether errors occur in the beginning or end of learning, which is key to our hypothesis.

6.2.2 Learning curves

If class-conditional independence underlies classification *early* but not late in learning, we expect that the critical items will be learned last frequent; specifically 111 is learned slower than 000 and followed by 001, 010, 100. A comparison of the simulated learning curves to the empirical learning curves for each of the five stimuli showed that the human learning curves equaled simulated learning curves but only for simulations with high values of the prior belief in class-conditional independence (above .7 for a parameter ranging from 0 to 1). Importantly, the most extreme parameter settings (a belief in class-conditional independence of zero or one) did not represent the obtained human learning curves. While these learning curves lend additional support to H3b, the aggregation of errors across participants ignores inter-individual variability.

6.2.3 Individual modeling

If class-conditional independence underlies the early stages of individual learning, we expect that for many people a model with high values of the prior belief in class-conditional independence parameter describes the development of choices best. The two model parameters were obtained jointly by one-trial-ahead predictions for individual learning trials using MSE as measure of data-model discrepancy. H3b predicts high values for the parameter governing the prior belief in class-conditional independence, π. In Experiment 1 the majority of our participants (27 of 30) showed high values of the prior belief ($\pi \geq .7$), only a minority (3 of 30) had no prior beliefs in class-conditional independence (i.e., $\pi = 0$ was the parameter best describing behavior). Similarly, in Experiment 2 most participants showed values of $\pi \geq .8$, only few participants (4 of 29) were best described by a fully flexible model with no beliefs in class-conditional independence. This corroborates the use of class-conditional independence early in learning.

6.2.4 Discussion and limitations

A robustly functional categorization system needs to address the complexities of the task. I hypothesized that the statistical principle of class-conditional independence, which is remarkably robust across different task structures, describes human behavior early in learning. That is under conditions of uncertainty when the task is unknown to participants. Later in learning, when humans gathered experience with the task, I hypothesized that they could give up their robust initial assumption and adapt to the environmental structure. In two experiments, I found that class-conditional independence describes individual and aggregate learning behavior. In line with the prediction, most peoples' initial decisions corresponded highly to class-conditional independence, but they were able to adapt to the

environment.

The current results emphasize the importance of studying the ecological rationality of cognition. Our findings are in line with previous research showing that in the early stages of category learning people employ a simpler categorization strategy, and then gradually learn more computationally intense strategies (J. D. Smith & Minda, 1998; Love, Medin, & Gureckis, 2004). Our work extends these findings by adding a notion of robustness to the notion of simplicity.

Regarding process models, I would like to be clear that the DISC-LM model is — according to the framework for process models — a model of behavior, not process. In its current version the model was designed to test a specific hypothesis about people's behavior, not so much about the cognitive processing underlying it.

6.3 Functional specification: Tracing the processes for evolved risk responses (Jarecki & Wilke)

Before addressing the hypothesis that risk propensities differ across domains, I wished to establish whether domains improved a model fit above gender, which is a major variable associated to risk taking (Byrnes et al., 1999). The data were more likely under the a model including the variables gender and domain as opposed to gender only ($AIC_{\text{full}} = 11,836$ vs. $AIC_{\text{gender}} = 13,066$), with a significant log likelihood difference, $\chi^2(18) = 1266.616$, $p < .001$. Therefore, subsequent analyses include both gender and domain in the models.

6.3.1 Domain differences

According to **H3a: The propensities to take risks differ across domains of various evolutionary functions, replicating previous findings**, risk propensities should differ across domains. The obtained effects — log odds of choosing the next higher likelihood category — show that risk taking propensities differed across domains and gender. Main effects[4] of domains were found for all domains except mate attraction and parental investment. For Status/Power and Mate Retention, people were least risk-taking. Women showed highest risk propensities regarding Food Selection and Kinship, whereas men showed highest propensities for Within-group Competition and Kinship risks.

Though in general women were less risk taking than men, it is important to note that in some domains women were *more* risk taking than men: regarding food selection, parental investment, and kinship women's mean interaction effect sizes compensated the negative main effect.

[4]Effect means a significantly different chance to choose the next higher category compared to the arbitrarily chosen baseline domain within-group competition.

6.3.2 Stability of domain differences

The current results replicate previous findings from student populations (study 2 and study 3 by Wilke et al., 2014). Eight of 10 main effects are similar between the present and the past studies (effects had at least the same direction in both studies, or were small/insignificant in both studies). Regarding the domain-gender interactions, all of the effects were replicated. These results provide converging evidence for a stable dependency of human risk taking propensities on evolutionarily motivated content domains.

The effects which the present study could not replicate concerned the domains of Food Selection and Mate Attraction. For Mate Attraction, the present sample was less risk prone. But our sample was also older and contained more married individuals than Wilke's study (see below); thus it is unsurprising from a life-history perspective that our participants were less prone to engage in mating risks.

The demographics of our sample differed from the the student sample of study 3 by Wilke et al. (2014). The present participants were older than the students (mean age 30 vs. 19 years), $t(123) = 13.987$, $p < .001$, Cohen's $d = 1.785$; the present sample also contained more parents (mean 0.61 vs. 0.06 children per person), $t(131) = 5.428$, $p < .001$, Cohen's $d = 0.680$; and more of our participants were married (54% vs. 22% married), $\chi^2(1) = 38.576$, $p < .001$, Cohen's $h = 0.676$. Similar differences hold in comparison to Wilke's (2014) study 2.

6.3.3 Cues for risk taking behavior

The next set of results addresses H3b through H3d.

Inter-rater reliability. The average inter-rater reliability, pooled across domains, was Gewt's AC = .69 (median .69, $SD \pm .05$, range .63 to .78). Given that coding was difficult because there were more than 30 categories per domain, the reliability can be considered fair to good agreement. In total, participants reported 1593 cues. People reported on average of 2.64 cues per question (median 2, $SD \pm 1.29$, range 1 to 9). The data show that the number of cues did not differ statistically across domains, one-way ANOVA ($number \sim domain$), $F(9) = 1.893$, $p = .113$, $\eta^2 = 0.460$. Bit note that the effect size is rather big, which indicates that there are differences undetected due to a lack of power.

Relative frequency of positive cues. Overall, about half of all reported cues were positive (pointing towards more risk-taking). The median number of positive cues was 1 in 3 for women, and 1 in 2 for men (the relative frequency was, on average, 47% and 48 %, respectively; with medians of 50% for men and women). Men and women reported equal proportions of positive and negative cues.

To address **H3a: (In domains with higher risk propensities, more positive cues than negative cues are retrieved)**, I analyzed the relative frequency of positive cues. If higher riskiness is associated with recalling more positive than negative cues, I expect that the pattern of the proportion of positive cues and the heights of risk taking propensities to look similar across domains. The results show that the cross-domain pattern of the relative frequency of positive cues did not follow the pattern of domain specific risk taking propensities, measured as the marginal effects of domain on risk taking from the model fitted before. This suggests that people do not rely on the retrieval of events which

favor risk-taking to determine their risk-taking propensities.

According to **H3b: In domains with higher risk propensities, more positive cues than negative cues are retrieved** I expected that the pattern of risk taking propensities is similar to the pattern of the order of cue directions across domains. The results show that the order of positive and negative cues in terms of the standardized median rank difference (SMRD) differs across domains, e.g. while for Between-group Competition negative cues were recalled prior to positive cues the reverse holds for Status/Power. However, the relative differences in the order of cue directions across domains showed little correspondence to the domain-differences in peoples' risk propensities. This suggests that retrieving positive earlier than negative events is not the driving force behind risk-taking.

According to **H3c: In domains with higher risk propensities, the specific cues are retrieved, independent of the cue direction**, I expected the recall of specific cues to follow a similar pattern to the cross-domain pattern of risk taking propensities. I obtained the three most frequently mentioned cues in each domain — independent of their direction — and computed how many of the men and women reported this one cue. The results show that for men, the 3rd most frequent cue follows a pattern rather similar to their risk taking propensities, except for the kinship domain. For women the pattern of the 2nd most frequent cue closely matches their risk taking propensities. This suggests that certain, gender-specific cues may underly domain-specificity in risk taking.

6.3.4 Discussion and limitations

With the present study I investigated the domain-specificity and variables related to the cognitive processes underlying risky choice. I hypothesized and found that risk taking propensities differ across functionally specified domains, and that these differences closely replicated previous findings. Further, I explored potential ways to use the retrieved cues and how they relate to the domain-differences using cues for risk from memory. The results show that the single, frequently recalled cues most closely related to the relative risk propensity in the domains. Further, our results show that the cue directions were not related to differences in risk propensities across domains.

These results emphasize the need to investigate the processes underlying preferential choices (choices without an immediate performance criterion), like the propensity to take or avoid risks. Our approach is complementary to process models that have investigated risk taking in discrete choice experiments using risky gambles (e.g. Brandstätter et al., 2006), however the existing models of risky choice tend to rely on monetary inputs (payoffs, probabilities). Our data can be used to specify a search rule for a decision heuristic: people search through the information about risk in different domains according to the retrieval frequency given in our data. This specification allows to test whether the utilization of the cues from memory that were identified, matches how people use these cues if they are presented as environmental variables.

I acknowledge that the results have an exploratory character because not every respondent reported cues for all domains. Therefore it was not possible to conduct fully individual-level analysis of the relationship of individual pattern of risk-taking and retrieved cues. I believe that further research needs to address whether people actually use the cues as described.

Methodologically, while questionnaire data is sometimes seen as less valid than choice data, recent findings (Frey, 2015, unpublished data) suggests that questionnaire measures are more reliable and consistent than measurements involving gambles. Further, several studies found risk questionnaires to correlate with frequencies of behaviors (Wilke et al., 2014). Hanoch, Johnson, and Wilke (2006) found that real-life risk takers, like bungee-jumpers, scored higher on the respective domains on the risk scale but not others. In sum this suggests that risk scales are valid approximations for risk behavior.

CHAPTER 7

GENERAL DISCUSSION AND FUTURE DIRECTIONS

The main theme of my dissertation was to progress the study of the form of cognitive processes based on their function. The integration of form and function from an evolutionary perspective was considered a fundamental question that decision making research should address (Hastie, 2001). The present thesis considered one theoretical obstacle toward the unification of form and function and applied two aspects of a functional analysis to risky choice and categorization.

My thesis first summarized a long history of investigations that either concern the function of decision behavior or the form of the process underlying it. The existing proposals to integrate the study of how the mind processes information and the purpose of this process (Bischof, 1998; Gigerenzer, Todd, & the ABC Research Group, 1999; Brunswik, 1955; J. R. Anderson, 1990) differ with respect to the models they suggest to describe decision making or the underlying process. The theoretical contribution of my thesis is a clearly specified framework for process models. The two most noteworthy properties are separability and testability, namely precise hypotheses for both the process and the resulting behavior in a way which avoids reverse inference. It has implementations for the design of models in cognitive science, and the types of models used in economics and evolutionary psychology.

The second contribution of my thesis concerns one of the most-studied fields of decision making: categorization. A functional cognitive system needs to perform robustly. In line with this I hypothesized and found that a robust principle from machine learning (class-conditional independence) describes peoples classification choices early in learning well.

The third contribution of my thesis concerns risky choice – another well-studied area. Specifically it concerns how cognitive processes relate to the evolutionary functions of risk taking in ten domains. It is the first to combine an evolutionarily-derived measure of risk taking with process tracing. Our results show that utilization of specific cues retrieved from memory is most closely related to the differences in risk propensities across domains.

I will now outline several implications of my findings for future research in the different fields related to human decision making that I reviewed in the introduction.

7.1 Direction for cognitive science

The present thesis opens up conceptual and methodological avenues for decision sciences. Methodologically, the process model framework provides a pathway for researchers to disentangle the decision

making behavior and the cognitive processing components within formal models. To date, many models in cognitive science do have parameters with psychological interpretations, as requested by Lewandowsky and Farrell (2010). Parameters tend to be fitted from behavioral data; for example attention weights in exemplar models (e.g., Nosofsky, Kruschke, & Mckinley, 1992). Unlike fitting process parameters, the process model framework outlined here advocates to specify these parameters from the input variables and use their values to derive testable hypotheses to predict process-level data. This idea is implicitly implemented in many fast and frugal heuristics (take-the-best, priority heuristic), but not all (e.g., 1/N, prospect theory). Separability, as a crucial property of models of process, is a qualitative property of a model that calls researchers to predict data across different levels of analysis.

The framework is further useful to both stir the discussion about process model requirements, and to avoid debates about the plausibility of a model as process model. Long-standing debates as to which model constitutes a proper process model (e.g., Pohl, 2011), need a definition of process model characteristics to be resolved. The framework is meant as a starting point for precisely characterizing, developing, and teaching how to build process models.

The findings regarding the domain specificity of risk taking have implications for several areas of cognitive science. The first implication is that including evolutionarily derived functional goals leads to greater stability in choice behavior. Our replication results of risk propensities suggest stable differences in risk taking propensities (at least for a North American population) and the differences we found are in line with evolutionarily relevant variables such as the number of offspring. This implies the following avenue for risky choice research: Complementary to studies on how to model and explain the heterogeneity of processes underlying risky choice (Pachur, Hertwig, Gigerenzer, & Brandstätter, 2013; Glöckner & Pachur, 2012), our findings suggest to identify and characterize evolutionary goals that imply homogeneous behavior, and progress investing processes from there. In this sense, form follows function.

The findings on risk also have implications for research on categorization. Sloman, Lombrozo, and Malt (2003) pointed out that categorization research tends to focus on two questions: on the one hand, there is work modeling the fine-grained information integration steps during categorization (e.g., Medin & Schaffer, 1978; Nosofsky, 1984). These studies usually present abstract stimuli, like rectangles or schematic faces (similar to risky gambles). On the other hand, there is work investigating how context changes category formation (e.g., Wattenmaker, Dewey, Murphy, & Medin, 1986). These studies present a broader variety of stimuli. The divide is almost identical to the one we observed in studies of human risky choice. The current findings from contextualizing the material to study risk taking showed that risk behavior is very stable within functionally derived content domains. This implies that categorization processes may also be more stable within evolutionary domains. This is a topic where biologists or anthropologists and cognitive scientists can meet: In biology, categorization is mostly studied as a similarity-based generalization process within specific domains, such as animal food classification or predator categorization. The cognitive literature is broader with respect to the categorization models used (similarity-based models, networks, decision trees), but when which strategy of classification describes human behavior best poses an unresolved and open question. Our replication results regarding risk taking imply for categorization, that combining biological functions

of category formation in evolutionary content domains with different cognitive category formation models yields progress for this question.

7.2 Directions for evolutionary psychology

Our findings provide several insights and directions for modeling in evolutionary psychology. As mentioned in the introduction, research here usually uses optimality-type models to determine the evolutionary stability of cognitive modules. In this field, much progress could be gained by shifting the focus away from simple a reductionist representations and toward slightly more complex but psychologically plausible process models, as outlined in the process model framework.

A second direction, implied by our results regarding categorization is the call for integrating robustness and dynamic adaptation. From an evolutionary perspective, categorization systems evolved for efficiently reducing information for action selection. For example, reducing all the features of potential predators to the class 'predator' enables to respond to the threat by selecting the appropriate action. Our results showed that people start out with a robust category computation, but are able to give it up. Within the evolutionary context, this opens up the question whether the combination of cues for predatory species are at variance with class-conditional independence, or if evolutionary environmental structures allowed class-conditional independence to robustly perform classification tasks.

The third direction consists of using process tracing to address hypothesis about the integration of cues which, according to evolutionary theory, are relevant factors for decision making and fitness. How cues are combined (e.g., in a compensatory or non-compensatory manner) is relevant for the adaptiveness of cognitive systems. Evolutionary psychology tends to consider which cues humans and animals attend to (Barrett & Kurzban, 2006). But how cues are integrated is another important aspect of the performance of the cognitive system. For example, evolutionary theories, such as risk-sensitive foraging (reviewed in McNamara & Houston, 1992), predict that the value of internal states like the own reproductive value, or the amount of resources should influence animal foraging decisions. Yet, how these variables are integrated to form a choice is less clear, and optimality models traditionally fail to make predictions regarding this. Aspect listing and other process tracing measures provide insights in the cue utilization, while process modeling provides the formal means to test whether a population of agents with a specific model is able to survive. The present results suggest that lexicographic processing is an important aspect that should be considered in work aiming to explain the processes underlying behavioral differences across fitness-relevant domains.

7.3 Directions for biology and ecology

For biology and ecology a methodological recommendation concerns the investigation of robustness through methods from machine learning. Our findings regarding class-conditional independence as robust strategy for early categorization choices suggest that mathematical results obtained in computer science and machine learning regarding the robustness of algorithms can be useful principles for the

investigation of human and animal choice as well.

7.4 Directions for behavioral economics

Behavioral economics would benefit from introducing process models according to the process model framework into their methodological portfolio. Besides broadening their methodological means of investigating market behavior, process modeling is particularly important because of the policy relevance of economics. Economic research often informs policy decisions. Laws and regulations are (at least partially) directed toward changing human behavior, e.g., criminal sentences serve to minimize crime rates, or higher taxation of fuel aims to encourage environmental friendly behavior. Insights into the cognitive processes underlying such choices are likely to result in policies with greater impact on decision makers. Knowledge of how information is integrated and to which degree certain cues shape choices provides a crucial tool for evidence-based policies.

7.5 Conclusion

To conclude, the present thesis advocates to stronger integrate the study of form and function of decision making. My thesis addresses three aspects of a form-function integration: Models of the form of decision processes, robustness, and evolutionary functioning.

Bringing these three aspects together yields the following research program: Start by defining a set of evolutionary goals a priori, analyze which cognitive mechanisms would robustly achieve such goals, and formalize how human and animal decision makers achieve them with process models with testable predictions about the processes.

ORIGINAL STUDIES

Jarecki, Tan, & Jenny (submitted)

A framework for cognitive process models†

Jana B. Jarecki*, Jolene H. Tan*, & Mirjam A. Jenny*

*Max Planck Institute for Human Development, Lentzeallee 94, D-14195 Berlin

Abstract

Our article proposes a general framework for cognitive process models. It offers guidance for the development of process models, specifies to what extent process models require a specific form (probabilistic form, algorithmic form), content (reaction time, neuronal processing), or data (process data, behavioral data). It provides dimensions on which process models can be compared, and constitutes a basis for a taxonomy of cognitive models. The theoretical framework proposed in this paper characterizes process models in general by four dimensions: (a) their scope is tri-modal individuating the information entering the cognitive system, the phenomena leading to the behavior of interest, and the behavior to be modeled; (b) they allow precise and testable predictions to be derived for the behavior and the process; (c) the process predictions can be derived separately from the behavioral predictions, and without reverse inference from the behavior, and (d) the information transformation in the model is plausible in the sense of being compatible with the contemporary body of knowledge about human cognition. The framework can be applied to cognitive models before or after they are empirically tested. Moreover, the framework can advance currently unresolved debates among scientists about which models merit the label.

1 Introduction

This article identifies shortcomings related to the current use of process models and proposes a framework to address them. So-called process models, models that aim to formalize the processing of information in the cognitive system, are among the most prominent and widespread means to study cognition. In the last decade, the term "process model" has appeared in roughly 12,400 documents from cognitive psychology; the citations of database-indexed papers using this term have increased steeply (even when controlling for a positive citation trend), see Figure 1; and there has been a corresponding growth in interest in process measures (Schulte-Mecklenbeck et al. 2011, p. 9). Figure 1 shows that in 2013 the articles mentioning process model were cited more than ones mentioning formal model and agent-based model.

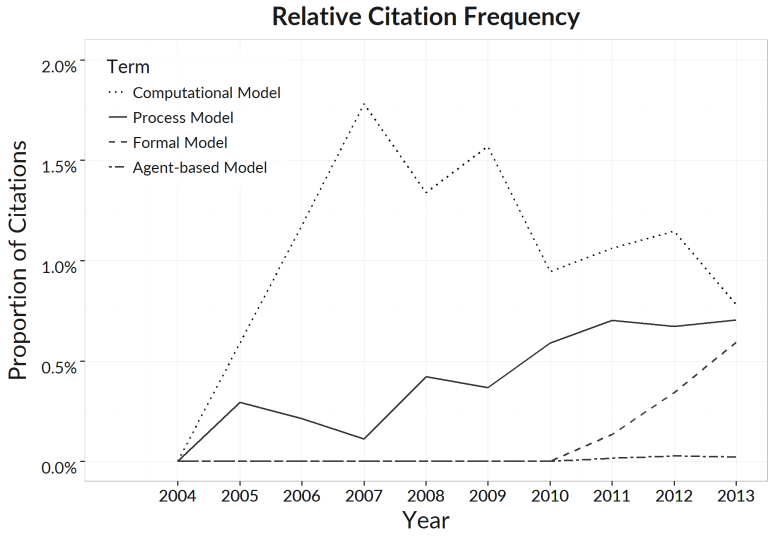

Figure 1. Increasing discussion of publications using the term "process model" with control for the rising numbers of citations. The solid line shows the proportion of citations of articles including the term "process model" AND cognitive science AND judgment and decision making (JDM) relative to citations including the latter terms but excluding "process model". The dotted lines depict the respective proportions for articles including the term "agent-based model", "formal model", or "computational model" in place of "process model." Cognitive science and JDM were operationalized as "cognitive," "psychology," AND "judgment and decision making" OR "decision making." Source: Web of Knowledge, accessed November 13, 2014. We refrained from a comparison with citations of specific mathematical methods, such as "Bayesian model," because it seemed unjustified to compare generic and specific search terms.

The shortcomings of this proliferation is that it is bottom-up in the sense that to date there is no framework which characterizes the properties of a generic process model. Textbooks (to date) offer limited instructions for process model development or the necessary characteristics to include in a process model. Lewandowsky and Farrell (2010) advise that parameters in a process model need a psychological interpretation (p. 18), and that process models need to describe the process in detail (p. 25). Despite this, we found little overarching guidance regarding which aspects matter for process

models. There is even less guidance about how to recast a model currently (arguably) *not* regarded as realistic model of the process into a process model. Disagreement as to which models constitute process models makes it infeasible to know which type of models are candidates. A further issue is that there is a limited connection between process models and process data. Process data is often theoretically required to test a process model (E. J. Johnson, Schulte-Mecklenbeck, & Willemsen, 2008; Weber & E. J. Johnson, 2009), but which data counts as such? For example, eye-movements may be process data (e.g, Lemonnier, Brémond, & Baccino, 2014; Orquin & Mueller Loose, 2013), or outcome data (e.g., Reichle, Rayner, & Pollatsek, 2003) depending on the research question. The next shortcoming concerns systematizing (process) models. We cluster models because they either model similar tasks (e.g., the general context model, RULEX, and prototype models are for classification; cumulative prospect theory, and the priority heuristic are for risky choice), or because they share formal features (e.g., drift-diffusion and accumulator models contain random walks). Yet, it is unclear which of those belong together as process models. The next limitation concerns model selection for model comparison. To date, models are (mostly) compared based on their behavioral predictions only, without regarding if they are supposed to describe processes. Lastly, in conferences and the literature we witness unresolved debates about whether a model rightfully merits the label process model.

The next section will present an empirical treatment and literature review of the disagreements related to process model properties. Then we introduce a four-dimensional conceptual framework for cognitive process models. Finally we illustrate how to apply it to one recent model of risky choice (Fischbacher et al., 2013), and conclude by discussing limitations.

1.1 State of the Art Regarding Process Models

The following section outlines the judgment and appearance of process models among experts and in the literature.

1.1.1 Expert Judgment

We surveyed scientists, among them many developers of cognitive models, asking them to classify 116 models according to whether they were process models. The list of models was derived from a systematic literature review (see Appendices A.1.1 to A.1.3 for details). Our sample consisted of 62 researchers, 35 professors, 16 post-doctoral researcher, and 11 doctoral students. Most had methods teaching experience ($N = 46$), and all were familiar with a good proportion of the 116 models: The professors, researchers, and students knew and classified on average 50, 49, and 40 models, respectively, indicating that the sample consisted of experts.

Although a high proportion of researchers (82% of 62 respondents) agreed that process models are important, they did not agree on which models constitute process models. We analyzed this terms of inter-rater agreement, measured by Fleiss-Cuzick's kappa[1]. We found $\varkappa = .27$, indicating low

[1]Which is a statistic of inter-rater reliability suitable for our data (i.e., dichotomous ratings by more than two judges with an unequal number of judges per item)

agreement.[2] A split by seniority yielded $\varkappa$ values of .33, .17, and .14 for professors, researchers, and students, respectively, implying that the low agreement was not merely an artifact of averaging over researchers' seniority levels.

This disagreement between judgments also suggests that the meta-theories related to process models, such as Marr's (1982) three levels of analysis (computation, algorithm, and implementation), have not provided a specific enough characterization of the properties of process models. The tri-level approach has been widely adopted (recently, e.g. Chater, 2009; Griffiths, Lieder, & Goodman, 2014; Griffiths, Vul, & Sanborn, 2012; K. Huang et al., 2012; Jones & Love, 2011; McClelland et al., 2010), but also criticized for being difficult to apply (summarized by Griffiths, Lieder, & Goodman, 2014). Process models tend to be located at Marr's algorithmic level, which is defined as specifying the algorithm that transforms an input into an output (Marr, 1982, p. 5). But Marr does not provide much more detail. This is also reflected in our survey responses from the 38 experts familiar with Marr: When asked whether the algorithmic level clarifies what process models are, their opinions were divided between "does not clarify at all" ($N = 16$) and "clarifies completely" ($N = 20$) around a "neutral" midpoint ($N = 2$) on a 7-point Likert-type scale.

1.1.2 Appearance of Process Models in the Literature

In the decision making literature, process models appear with different connotations. Do these connotations converge towards a shared understanding of process model properties? The next section reviews the context and circumstances in which process models appear. Note, that we wish to not imply that any of the authors meant to define process models explicitly through their use of the term.

Context 1. Process Models Compared to Rational Models. The class of rational models (e.g., Chater, 2009; Lewis, Howes, & Singh, 2014; Sanborn et al., 2010) provide the ideal solution to the abstract statistical problems that organisms have to solve (Griffiths, Vul, & Sanborn, 2012); they are related to J. R. Anderson's (1991a) rational analysis, that aims to model the optimal behavior from entering people's goals and capacities into a formal model of the environment. In part of the literature, process models appeared in contrast or competition to rational models (Griffiths, Vul, & Sanborn, 2012; Chater, 2009; Jekel, Glöckner, Fiedler, & Bröder, 2012). For example, Lee and Cummins (2004) introduce their paper comparing rational and process models (similarly Bergert & Nosofsky, 2007). Given this, a scholar could build a notion of process models as models yielding solutions that are *not* guaranteed to be optimal or that have proven to be suboptimal, or as models yielding approximately optimal solutions within a fixed margin of error (so-called rational process models; see Griffiths, Vul, & Sanborn, 2012; Griffiths, Lieder, & Goodman, 2014; Sanborn et al., 2010). Accordingly, a process model developed by somebody with these connotations qualifies itself by a *suboptimality criterion*: a model at most near-optimal or even suboptimal choices.

Context 2. Process Models Compared to As-if Models. Secondly, let us consider as-if models (e.g., Berg & Gigerenzer, 2010; E. J. Johnson, Schulte-Mecklenbeck, & Willemsen, 2008). As-if

[2] $\varkappa = 0$ indicates random agreement; $\varkappa = 1$ indicates perfect agreement; values above .60 are considered to indicate "good" agreement (Fleiss & Cuzick, 1979).

models involve input-output transformations that do not correspond to factual phenomena in the modeled system (see, Glöckner & Witteman, 2010). Their mathematical representations are chosen for elegance or feasibility and deliberately free from psychological interpretations; they describe behavior, but not the processes (Brandstätter et al., 2006). As-if models relate to Milton Friedman's 1953 arguments for positive economics (which is often cited as favoring as-if models). In the literature, process models appear as distinct from such (Gigerenzer, Todd, & the ABC Research Group, 1999; Sanborn, 2014). Chase, Hertwig, and Gigerenzer, 1998, for instance, contrasted models assuming unlimited computational resources with models assuming computational constraints. From this second context, one may conclude that process models are characterized by a *feasibility criterion*: their computations are required to be feasible given human mental capacities, or their parameters need a psychological interpretation (e.g. Gigerenzer, Todd, & the ABC Research Group, 1999; Goldstein & Gigerenzer, 2002; J. I. Myung, Pitt, & Kim, 2003).

Oftentimes, rational models (that yield optimal solutions) are seen as-if models (with unrealistic computation). This is because computing optimization routines would require unlimited mental capacities. Accordingly, process models are neither rational nor as-if models, as Gigerenzer and Goldstein (1996) write. Although rational and as-if models can be identical, they need not. Models that are not as-if (involve realistic computations) can outperform unrealistic models with limited data (Gigerenzer, Todd, & the ABC Research Group, 1999). Further whether a realistic model is optimal, hinges on the criteria for optimality (Chase et al., 1998; Einhorn & Hogarth, 1981; Marcus & Davis, 2015).

Context 3. Process Models' Share Formal Features. In a different context, the term process model occurs in the relation to formal aspects of modeling. Process models have been related to stochastic computations, like random walk processes (in drift-diffusion or in accumulator models Busemeyer & Townsend, 1993; Ratcliff, 1978; Brown & Heathcote, 2008; Pike, 1973). Process models have also been related to specifically developed symbolic languages (Einhorn, Kleinmuntz, & Kleinmuntz, 1979; Gregg & H. Simon, 1967; H. A. Simon & Kotovsky, 1963) like Newell's 1963 Information Processing Language-V. In this context, process models are characterized by a set of *formal properties* related to the modeling paradigm.

In sum, the contexts in which process models appear in different parts of the cognitive literature, are distinct from each other. The first stresses suboptimal performance of the decisions predicted by process models. The second context stresses feasibility of the computations implemented in the model. The third context stresses formal elements, such as stochasticity. This brief review corroborates the earlier empirical findings that the key characteristics that constitute cognitive process models are unclear.

1.1.3 Summary

To sum it up, researchers' opinions diverge concerning which models constitute process models, and the literature uses the term in a diverse set of contexts. At present it seems hard to teach process modeling without invoking one of the reviewed connotations. We will introduce a general framework

for process models below and incorporates part of the contexts just reviewed.

2 A Framework for Process Models

Our framework can be applied to cognitive models when they are constructed—that is, before they are tested—as well as after they are tested. We will illustrate it using cognitive models, and hypothetical models of a cash register from Mars, that cannot be opened; because this example illustrates well and requires no modelling knowledge. Importantly, the framework refers to empirical models of human behavior that purport to describe cognitive processing.

In the following, we refer to the processed information as input and to the resulting behavior of interest as output. According to our framework in Figure 1 process models include one or more intermediate stages, are built such that the intermediate stages are compatible with current knowledge about human cognition, construct the intermediate stages such that they can vary separately from the output, and allows testable claims to be derived about both the intermediate stages and the output, given the same input.

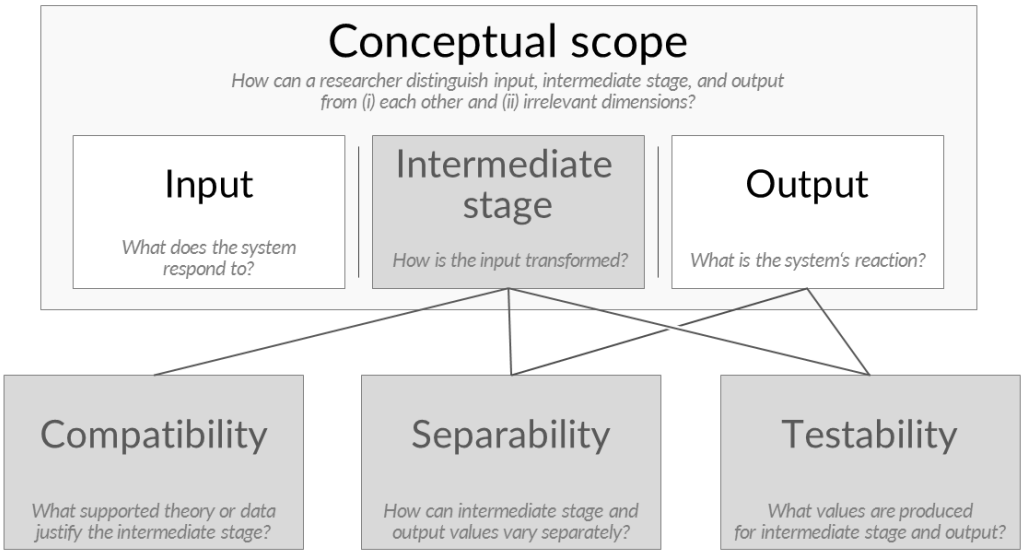

Figure 1. A framework for cognitive process models. The schema shows the requirements for a model of cognition (conceptual scope, input, and output) as well as the additional requirements a model should fulfill to qualify as a process model (intermediate stage, compatibility, separability, and testability). The connecting lines denote the interrelatedness of the process-model specific requirements in the gray boxes.

2.1 Tri-Modal Conceptual Scope

A model's scope describes the phenomena to which it applies. For non-proces models the scope includes only two aspects. Those are the input phenomena and the output phenonema. For example, consider

two models of behavior in economic games. The fairness models by both Bolton and Ockenfels (2000) and Fehr and Schmidt (2010) describe peoples' contributions to the game given social preferences and the value of relative payoffs. The social preference in those models is defined in terms of the output (as the difference between a person's payoff and the average of everyone else's payoff). Thus, the model includes only inputs (payoff structures) and outputs (behavior). It has a bi-modal scope. In process models, the scope ought to be tri-modal. A tri-modal scope describes the phenomena related to inputs, process, and outputs. Moreover, a tri-modal scope individuates the model's input and output from the process explicitly. For example, Gluth, Rieskamp, and Buchel (2012) offer a tri-modal scope when testing a sequential sampling model of decisions to acquire goods. They individuate the input (sequential information about a the good's value) from the process (the neuronal firing underlying the evidence accumulation) and the output (reject or accept decisions).

Note that a complete specification of the scope of a cognitive model (foreseeing all possible interpretations of what it might model), is infeasible.

2.2 At least one intermediate stage

Our framework requires process models to have at least one intermediate stage specified (in line with Svenson, 1979; Weber & E. J. Johnson, 2009). An intermediate stage is the event or the events that—according to the model—occur after the input and before the output (see Figure 2). It can be latent or manifest, continuous or distinct, but must lie within the conceptual scope. Intermediate stages could be, for instance, where a person is looking or looks next; which brain regions are activated; beliefs about probabilities; or causal structures; or which information a person compares. In specifying the requirement of intermediate stages, we try to be more precise than Marr's (1982) algorithmic level. The algorithmic level contains an information "transformation", which is rather broad.

For example, cumulative prospect theory (Tversky & Kahneman, 1992) lacks an intermediate stage. Prospect theory describes risky choice as computation of an utility $u(x, p)$ from multiplying people's subjective evaluation of the outcome $v(x)$ and their weight of the probability $w(p)$. The initial prospect theory proposed that people engage in probability weighting before outcome evaluation (Kahneman & Tversky, 1979). But the formal model contains no temporal order. The equation $u(x, p) = v(x)w(p)$ leaves open whether the mind first weights probabilities and then evaluates payoffs. This exemplifies a model without clearly specified intermediate stages. The fact that (this version of) prospect theory lacks an intermediate stage is independent of the question whether the mind really carries out the multiplication, which is an empirical question.

It is not the inclusion of equations in a model that precludes the existence of an intermediate stage, but a lack of temporal succession. Decision field theory (Busemeyer & Townsend, 1993) includes several equations—for instance, a random walk that determines how the propensity to choose one option over the other develops over time. These equations, however, describe a sequential sampling of information. This constitutes clear intermediate stages.

2.3 Compatibility

The information transformation in the intermediate stage of a process model should be compatible with our knowledge of cognitive capacities. Compatibility means relating the hypothesized process (a) to at least one supported theory or (b) to data about the capabilities of the system. This can be a, for example, a theoretical argument (a model's intermediate stage computations are not intractable), an empirical argument (a model's process does not exceed memory limitations), or a reference to data (a model's process is in line with empirical phenomena).

Compatibility specifies the call for plausible mental processes in cognitive models. For example, Winkel et al. (2014) required process models to be plausible. The exact meaning of plausibility, however, tends to be subjective (for a potential list of criteria Gigerenzer, Hoffrage, & Goldstein, 2008). Compatibility is objective plausibility: Our criteria for compatibility can be verified by other researchers.

These examples illustrate compatibility. Busemeyer and Townsend (1993) linked the computations in decision field theory to findings from approach-avoidance research, and to choice response-time theories. Similarly, the process hypothesized in the priority heuristic Brandstätter et al. (2006) assumes that people prefer a gamble if its lower payoff exceeds that of the other gamble by at least 10%, where the threshold of 10% is justified by reference to the culturally embedded number system (i.e., is rooted in an existing theory).

Why compatibility? There are two reasons for this criterion. First, the latent nature of cognitive processing warrants a theoretical or empirical justification. Second, the intermediate stage or stages are under-determined by input—output relations (Moore, 1997). If multiple possible intermediate stages lead to the same connection of inputs to outputs, one way to distinguish them is to provide a theoretical argument for greater compatibility of one intermediate stage.

2.4 Separability

Separability is maybe the most important aspect of process models. It means that the intermediate stage or stages and the output explain or predict two separable dimensions within the conceptual scope. This means the model allows to derive two separate predictions for the two dimensions. Predicting separable dimensions means that a model can predict output values correctly while predicting intermediate stage values incorrectly, and vice versa. Without separability, each correctly predicted output implies a correct process. Figure 3 illustrates that if the separability condition is not fulfilled, only part of the hypothesis space can be supported by data (in the shaded cells), but separability allows more fruitful tests. Separability avoids equating support for an output prediction with support for hypothesized processes, the logical fallacy of affirming the consequent (Geis & Zwicky, 2011, p. 562) Such equating is unproblematic for output models, but for process models because the processes are under-determined by input—output relations.

Separability can be seen as a prerequisite for model-based process tracing. A model of the form *input* + *attention* → *output* renders eye-tracking data a proxy for attention and therefore process data, whereas a model of the form *input* + *brain activity* → *attention* renders eye-tracking data a measure of the

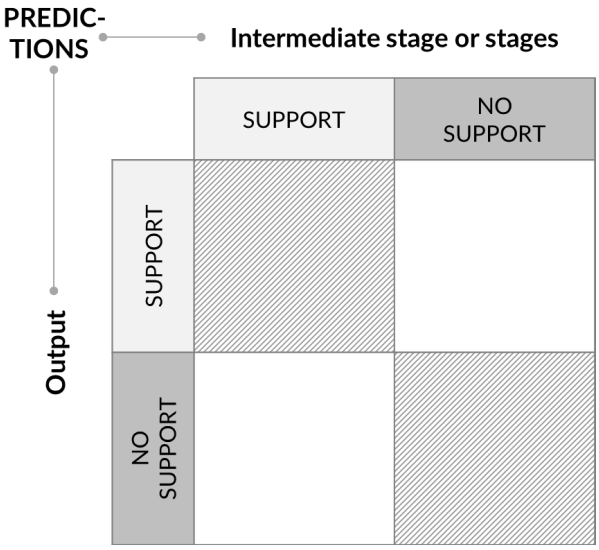

Figure 2. Implications of separability for model testing. Support and *No Support*: the data support or fail to support a prediction. *Predictions—Output*: values from the model at the outcome level *Prediction—Intermediate stage*: values from the model at the process level. For details, see text.

outcome. Models that fulfill separability provide the necessary frame that renders data, irrespective of whether it is eye-tracking or neuroimaging data, as process data, because these models predict at least one behavioral and one process-dimension.

Further, separability implies that a model in which all parameters are free is an outcome model. If a linear regression model (e.g., Dawes & Corrigan, 1974) is used to describe choices, and the weights are fitted from the choices, there is no separability. Separability implies that process parameters of process models are not by definition re-inferred from outputs. Note that having free or fixed parameters is not per se decisive for process models; what counts is separability.

Why separability? The reason for this criterion is model refinement. The predictions of random walk models, for instance, refer to choice distributions and reaction time distributions. These are separate dimensions. Early random walk models predicted response times for correct choices (summarized in Ratcliff & Tuerlinckx, 2002), but the data typically showed faster response time for errors than for correct choices (Ratcliff & P. L. Smith, 2004; Ratcliff, Van Zandt, & McKoon, 1999). If the researchers had not been able to test response time predictions separately, but had just fitted them, this discrepancy might have gone unnoticed. Instead, research was able to falsify the process component of earlier models, leading to their refinement. Thus, separability allows for empirical tests of processes, instead of mere assumptions.

2.5 Testability

The next criterion is testability. It refers to that the model's claims for the output and the intermediate stages can be tested on data within the scope. Figure 2 shows that testability refers to output and

process. Output testability is nothing special for empirical models, but process models need to yield additional testable hypotheses for any intermediate stages encompassed by the scope. The psychological constructs in the intermediate stages should be specified precisely so that it could be operationalized and tested by other researchers.

To sum it up, if a model contains at least one psychologically-motivated intermediate stage temporally located between input and output, it is a candidate process model. Importantly, cognitive process models need to yield separate hypothesis or predictions for the processes and for the behavior. This allows the two levels to be disentangled empirically, while they remain connected through the model and theory. In presenting the model, scientists need to be clear about which behavior, processes, and inputs it refers to by using a tri-model conceptual scope. And the hypothesized process ought to be compatible with current scientific knowledge of mental capacities (within the scope of the model).

The above is not meant to value process models over outcome models. It aims at defining the term. In discussion with colleagues and at conferences, "process models" tends to be used like "good model". However, which model is good depends on its purpose and explanatory power, as well as the research question.

2.6 Interrelatedness of the Components of the Framework

The connecting lines in Figure 2 illustrate the dependencies among the process model framework's elements. Three dependencies are noteworthy.

First, the intermediate stage is a prerequisite for the other characteristics (compatibility, separability, testability), that all apply to the intermediate stage. We believe intermediate stages are uncontroversial process model properties. This means that models such as cumulative prospect theory (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992) and weighted additive models do not, at least in their simplest form, qualify as process models in our framework. Second, the separability element is connected to the intermediate stage and the output. This is because the output prediction needs to be separated from the intermediate stage prediction. Third, testability connects to the intermediate stage and the output, because a process model—as opposed to a model of behavior only—requires predictions for both, behavior and intermediate stage.

In addition to these dependencies, the conceptual scope constrains testability. Models can only be tested within their scope. For example, one model about how people decide in a mini-ultimatum game (a decision tree by Fischbacher et al., 2013) is a lexicographic model that proposes sequential processing of information as intermediate stages. The scope of this model covers behavioral choices and visual processing, but not the exact firing of neurons in the prefrontal or visual cortex. Accordingly, the testability of the fast and frugal tree refers to testable hypotheses about the order in which people attend to information and not about the neuronal processes involved. Another example: To predict binary choices, the take-the-best model (Gigerenzer, Todd, & the ABC Research Group, 1999; Gigerenzer, 1991) proposes as intermediate stage that informational cues are ordered from the most to the least valid cue. It does not explicitly model how the cognitive system arrives at the validity ordering. According to the authors, this lies beyond the conceptual scope—and according to our framework, the testability

of the take-the-best model therefore cannot refer to how the humans estimate validities. A different or extension of this model is needed for the process of validity estimation.

3 Application Example

We will illustrate the framework by reference to the work of Fischbacher et al. (2013), who proposed a fast and frugal tree model of a responder's decisions in a mini-ultimatum game . Let us first briefly describe their model. It models the decision to accept or reject an offer in the ultimatum game. The ultimatum game is an economic experiment with usually two experiments. One participant is assigned the role of a proposer. The proposer is endowed with points or money, and decides whether to share part of it with the other participant. The receiver then decides whether to accept or reject the allocation. If she rejects, both participants receive nothing, if she accepts the allocation is implemented.

The fast and frugal tree model describes individual decisions in the ultimatum game in terms of three discrete steps (Fischbacher et al., 2013, p. 466):

— Step 1. If a responder receives more than the proposer, the responder will accept the offer without further ado.

— Step 2. If the responder receives less than the proposer, the responder will consider whether the offer was nevertheless kind and—if it proves kind—will accept it.

— Step 3. If the offer is considered unkind, the responder will ask him- or herself: "Would I have made this offer?" If the answer is yes, the responder will accept the proposed allocation; otherwise, the responder will reject it.

Guided by the process model framework, we next discuss whether this model provides a clear conceptual scope and how well it complies with each of the four characteristics of a process model: intermediate stage, compatibility, separability, and testability.

Tri-modal Conceptual Scope. A tri-modal conceptual scope is given if the authors define not only the input and the output, but also the intermediate stage or stages and the levels at which they are situated. The output in the model is the responder's behavioral decision in the game: to either accept or reject the proposer's offer. The input is the offer. More specifically, the inputs are four binary attributes of the offer and their respective values: Is the responder's payoff positive? Is the responder's payoff at least as large as the proposer's? Is the proposer's offer kind? Is the proposer's offer what the responder would have made? The intermediate stage is the order of information processing which, in the model, corresponds to a lexicographic consideration of the cues. The tri-modal conceptual scope is thus clear.

Intermediate Stage. A process model requires at least one intermediate stage in terms of an event between input and output. The model specifies four distinct mental events, as described above: thinking about the payoff, thinking about the relative magnitude of the payoff, thinking about kindness, and comparing the offer to one's own potential offer. The model thus includes more than one intermediate stage.

Separability. Separability holds if, given some input values, the values predicted for the output and the

values predicted for the intermediate stages can vary independently without reverse-inference. In the Fischbacher et al. (2013) model, the data reflecting the intermediate stage or stages are reaction time data. They can vary independently of the data that reflect the output, which are decision behavior data. Accordingly, the model might predict the decisions correctly, but not the reaction times (or vice versa). Measurement separability is therefore also given.

Testability. Model claims are testable if the model allows specific predictions to be made for the output and intermediate stage or stages, such that data can contradict them. The specification of the Fischbacher et al. (2013) model allows two predictions to be made. Specifically, the model predicts (i) the responder's behavior in the ultimatum game: will she reject or accept? (output prediction) and (ii) that the more cues the responder checks before deciding, the longer the response latency will be (intermediate stage prediction). Both predictions are precise and lie within the scope of the model. Therefore, testability is also given.

Compatibility. The model is theoretically possible if the intermediate stage or stages are argumentatively linked to either a supported theory or to data related to the conceptual scope. Fischbacher et al.'s (2013) argument for the compatibility of the intermediate stage based on previous support for bounded rational heuristic models, of which fast and frugal trees are a subclass. Therefore, a link to supported theories is given.

In sum, we have shown that Fischbacher et al.'s (2013) fast and frugal tree model offers a clear conceptual scope and possesses each of the four characteristics of a process model outlined above. Thus, drawing on our framework, we conclude that the Fischbacher et al. model is constructed as a process model.

4 Discussion

In the present article we proposed a general process model framework. It conceptualizes cognitive process models as descriptive models with multiple interrelated properties. That is, a model that includes at least one intermediate stage between input and output (this is a minimum requirement for a process in a process model). Further, it needs a clear conceptual scope spanning not only the input and output, but also the intermediate stage or stages. It needs to be set up such that it provides testable hypotheses within its scope not only for outputs but also for the intermediate stage or stages. Moreover, a process model needs to be flexible enough to let the data reflecting the output vary independently of the data reflecting the intermediate stage or stages (this is a minimum requirement for process falsifiability and tests using process data). Finally, the non-observable intermediate stage or stages need(s) to be compatible with current knowledge about cognition within model's scope, but need not be at the same level of abstraction as the intermediate stage or stages.

Conceptual clarity and consensus about the meaning of frequently used terms are desirable in their own right. If consensus is elusive, it is at least desirable to stimulate discussion about a term's meaning. Further, the framework can provide guidance to researchers investigating a novel phenomenon or task who seek to build a formal process model thereof. It may also guide those seeking to transform a model that describes only output into a process model. For example, it may guide researchers

wishing to construct cognitive process models of behavior in interactive game theory situations, which are frequently modeled by utility calculations (Fehr & Schmidt, 1999) in behavioral economics. Additionally, due to separability and testability, models built and reported according to our framework, make it easy to distinguish process data from non-process data. Consequently, it eases bridging the to date unfortunately separated fields of process tracing research and cognitive modeling.

In the same manner, models fulfilling our compatibility criterion (through being related to a supported theory or data) add a more objective notion to the notion that process models need to be "plausible". As a side-note, connecting to other theories is one way to foster theory integration. If the connected theory is questioned, as science progresses, this then automatically impacts the process model's status. Which models constitute process models is therefore, truly dynamic. Another benefit is that, according to our framework, process models necessarily describe data at various levels of analysis. They thus offer not only more opportunities to be tested but also more opportunities for model integration. Process models could be integrated not only horizontally but vertically—for example, by modeling choice data using process models about the order of perception, modeling the order of perception using process models of visual neuroscience, and modeling neuronal responses using process models of neuronal integration.

The rationale for including these four, and not other, process model characteristics is that they are independent of the formal notation (stochastic vs. deterministic, verbal vs. statistical, parallel vs. serial, etc.). Secondly, the characteristics force the model to link to data (testability, separability, and compatibility indirectly or directly link the model to data). Thirdly, our framework ignores matters of optimality and sub-optimality detailed in the introduction, for this is an entirely different question related to the choice of optimality criteria rather than models. We obviously did want to include model criteria derived from a theory, such as the search-, stopping- and decision rules presented by the fast and frugal heuristics program (Gigerenzer, Todd, & the ABC Research Group, 1999).

Our framework process *and* outcome, predictions. Why not test process models just regarding to the behavioral outcome? Because tests on the behavioral level suffice for input-output models. If one wants to compare process models, the process need to testable. Imagine researchers find that one model predicts the amount to pay by the Martian cash register better than another model. They then conclude the process proposed by the better model is likely the way the cash register actually works. But the only thing they can conclude is that the worse model proposed is unlikely to be the way the cash register works. Otherwise, they draw strong conclusions about something they did not even test, affirming the consequent (a common logical fallacy, Geis & Zwicky, 2011). Our framework gives researchers complementary dimensions on which models can be compared and evaluated as to whether or not they are process models.

4.1 Advancing Ongoing Process Model Debates

The process model framework has implications for ongoing debates about process modeling. There are two debates.

The first debate is normative. It asks whether process models are more useful than other models. The

question goes back to the tension between behaviorism and cognitivism, but the current debate focuses on models, rather than the legitimate content of theory. Some scholars argue that process models provide more realistic models of the mind than do structural models of judgment outcomes (Svenson, 1979) or economic as-if models (Berg & Gigerenzer, 2010; Gigerenzer, Todd, & the ABC Research Group, 1999). In contrast, others claim that rational models describe the mind better than process or mechanistic models (Chater, 2009).

The second debate is about model classification. The literature is filled with debates about what counts as a process model of choices (literally Brandstätter et al., 2006, p. 427). For example, do connectionist network models describe processes (McClelland et al., 2010) or functions (Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010)? Is the recognition heuristic (Goldstein & Gigerenzer, 2002) a process model (Pohl, 2011)? Do quantum probability models provide insights into cognitive processes (Pothos, Busemeyer, & Trueblood, 2013)? What process data can be predicted by the priority heuristic (Ayal & Hochman, 2009; Brandstätter et al., 2006; E. J. Johnson, Schulte-Mecklenbeck, & Willemsen, 2008)? Does decision field theory (Busemeyer & Townsend, 1993) predict the process of information search (Glöckner & Herbold, 2011)?

These debates—which models constitute process models, and which are more useful to describe the mind—are unresolvable without a general process model framework.

4.2 Conclusion

Our framework provides qualitative dimensions to guide the development of process models, frame the comparison of models in different languages, and foster conceptual clarity about process models. In most cases, models are compared by reference to model fit or prediction metrics. The metrics rank order models in terms of goodness of fit, or select one model based on critical tasks in which multiple models make different predictions, such that the data and statistical methods can disentangle the models. Both of these methods—obviously indispensable in empirical research—rely on data. For conceptual comparison, and eventually for conceptual theory integration, a general framework is necessary.

It offers a basis for a taxonomy of cognitive models and may increase consensus about which models can be considered process models. Importantly, the process-model specific elements and their interrelations provide a checklist to help scholars build new cognitive process models. The generality of our framework makes it usable to teach students about the concept of a process model. And finally it can advance currently unresolved debates about model types. Hopefully, this article further stimulates discussion about the term process models, increases the current discussions' structure and supports a more considerate and exact use of the term "process model".

Jarecki, Meder, & Nelson (2013)

The assumption of class-conditional independence in human category learning[†]

Jana B. Jarecki[*], Björn Meder[*], Jonathan D. Nelson[*]

[†]Jarecki, J., Meder, B., & Nelson, J. D. (2013). The assumption of class-conditional independence in category learning. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Cooperative minds: Social interaction and group dynamics. Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 2650-2655). Austin, TX: Cognitive Science Society.
[*]Max Planck Institute for Human Development, Lentzeallee 94, D-14195 Berlin

Abstract

This paper investigates the role of the assumption of class-conditional independence of object features in human classification learning. This assumption holds that object feature values are statistically independent of each other, given knowledge of the object's true category. Treating features as class-conditionally independent can in many situations substantially facilitate learning and categorization even if the assumption is not perfectly true. Using optimal experimental design principles, we designed a task to test whether people have this default assumption during categorization and learning. Results provide some supporting evidence, although the data are mixed. What is abundantly clear is that classification behavior adapts to the structure of the environment: people quickly learn a category structure that is unlearnable under the assumption of class-conditional independence.

1 Introduction

Categorization is fundamental for cognition. Grouping together objects or events helps us to efficiently encode environmental patterns, make inferences about unobserved properties of novel instances, and make decisions. Without categorization we could not see the woods for the trees.

Despite the ease with which we form categories and use them to make inferences or judgments, from a computational perspective categorization is a challenging problem. For instance, different diseases can cause similar symptoms, entailing that diagnostic inferences are often only probabilistic. Patients may have new symptom combinations and still require a diagnosis. Depending on the specific assumptions the physician makes about the relationship between the diseases and symptoms, a physician could justifiably make very different inferences about the diseases.

In the present paper, we investigate the role of the possible assumption of *class-conditional independence* of features in categorization. Class-conditional independence holds if the features of the category members are statistically independent given the class. This assumption can facilitate classification and learning of category structures. The concept of class-conditional independence is well known in machine learning, where it underlies the naïve Bayes classifier (Domingos & Pazzani, 1997) and is also a key assumption in some psychological classification models (Fried & Holyoak, 1984; J. R. Anderson, 1991b). It is related to ideas of channel separability in sensory perception (Movellan & McClelland, 2001). Similar ideas are found in Reichenbach's (1956) common-cause principle in the philosophy of science and in causal modeling (Spirtes, Glymour, & Scheines, 1993; Pearl, 2000). Both the philosophical and psychological literature make claims about the normative bases of the assumption of class conditional-independence of features. Our focus here is not on the general normativity or nonnormativity of that assumption, but on whether the assumption of class-conditional independence may (perhaps tacitly) underlie people's inferences in learning and multiple-cue categorization tasks. We think of this assumption as one of many possible default (heuristic or meta-heuristic) assumptions that, if close enough to an environment's actual structure, may facilitate learning and inferences.

2 The Psychology of Conditional Independence

Some psychological models of categorization incorporate assumptions of class-conditional independence, such as the category density model (Fried & Holyoak, 1984) or Anderson's (1991b) rational model of categorization. Both models treat features of instances as class-conditionally independent to make inferences about category membership or unobserved item properties. Other research has focused more directly on the role of conditional independence assumptions in human reasoning. For instance, a key assumption in many formal causal modeling approaches (e.g., Pearl, 2000; Spirtes et al., 1993) is the *so-called causal Markov condition*, which assumes that a variable in a causal network is independent of all other variables (except for its causal descendants), conditional on its direct causes. As this assumption facilitates probabilistic inferences across complex causal networks it was suggested that people's causal inferences could also comply with this conditional independence assumption. von Sydow, Meder, and Hagmayer (2009) investigated reasoning about causal chains and found that

subjects' inferences indicated a use of conditional independence assumptions, even if the learning data suggested otherwise.[3] Other research, however, found violations of the causal Markov condition (Rehder & Burnett, 2005). Asked to infer the probability for one effect, people's judgments were influenced by the status of the other effects rather than treating all effects as independent of each other given the cause. One explanation for this "nonindependence effect" (Rehder & A. B. Hoffman, 2005, p. 274) is that it might be due to subjective explanations that disable all causal links between the cause and effects at once (Walsh & Sloman, 2008). Other researchers have argued that these Markov violations do not indicate flawed human reasoning, but reflect the use of abstract causal knowledge that is sensitive to contextual information (Mayrhofer, Hagmayer, & Waldmann, 2010).

3 Research Questions

Should the assumption of class-conditional feature independence be used in classification learning? Do people use that assumption to guide learning about the structure of a novel environment? We extend previous research fourfold: (1) We use optimal experimental design principles (J. I. Myung & Pitt, 2009; Nelson, 2005) to explicitly address the assumption in classification, (2) we are interested in categorization learning as opposed to causal reasoning, (3) we investigate how people's experience with a new environment shapes their classification behavior, whereas many previous studies have measured explicit numerical probability judgments. (4) We use an experience-based research paradigm, whereas previous studies used numerical (Rehder & Burnett, 2005) or verbal (Mayrhofer, Hagmayer, & Waldmann, 2010) formats. Personal experience of events has been shown to result in different behavior and learning than word- or number-based presentation of probabilities (Hertwig, Barron, Weber, & Erev, 2004; Nelson, McKenzie, Cottrell, & Sejnowski, 2010). Before describing the task we designed, let us turn to the normative question of class-conditional independence in classification.

3.1 Class-conditional Independence in Classification

Categorization entails assigning an object to a class. Let F denote an object consisting of a vector of feature values $\boldsymbol{f}$, and let C denote a random variable whose values are the possible classes $c_1, \ldots, c_n$. The posterior probability of the class given the observed feature values, $P(\text{class} \mid \text{features})$, can be inferred using Bayes' rule:

$$P(C = c \mid F = \boldsymbol{f}) = \frac{P(F = \boldsymbol{f} \mid C = c)P(C = c)}{P(F = \boldsymbol{f})} \tag{1}$$

where $P(F = \boldsymbol{f} \mid C = c)$ denotes the likelihood of feature value vector $\boldsymbol{f}$ given class c, $P(C = c)$ is the prior probability of the class, and $P(F = \boldsymbol{f})$ is the occurrence probability of the feature configuration. An important question is how we estimate the relevant probabilities to infer the posterior probability. Estimating the classes' prior probabilities, $P(C = c)$, from the data is relatively straightforward. However, estimating the likelihood of the features given the class, $P(F = \boldsymbol{f} \mid C = c)$,

[3] For instance, applying the causal Markov condition to a causal chain $X \rightarrow Y \rightarrow Z$ entails that Z is independent of X given Y (e.g., $P(z|y,x) = P(z|y,\neg x)$).

is more complicated, as the number of probabilities grows exponentially with the number of features, feature values, and classes (the curse of dimensionality). Also, in many real-world situations there may simply be not enough (if any) data to derive reliable estimates for all possible feature configurations. One way to sidestep the problem is to assume that features are class-conditionally independent.

3.2 Class-Conditional Independence

If class-conditional independence holds the individual features within a class are statistically independent (e.g. Domingos & Pazzani, 1997). This means that the probability of a feature configuration given a class can be factorized such that:

$$P(F = \boldsymbol{f} \mid C = c) = \prod_{j=1}^{J} P(F_j = \boldsymbol{f}_j \mid C = c) \tag{2}$$

where $P(F = \boldsymbol{f} \mid C = c)$ denotes the likelihood of the feature configuration given the class, $P(F_j = \boldsymbol{f}_j \mid C = c)$ is the marginal likelihood of the j^{th} feature value given the class, and $j = 1, \ldots, J$ indexes the different features. Thus, according to the assumption of class-conditional independence, the likelihood of each feature value combination can be estimated from the likelihoods of the individual feature values.

3.2.1 Advantages

The key advantage of assuming that features are class-conditionally independent is that it reduces the curse of dimensionality. For example, for 10 binary features there are 210 possible feature configurations. That means, we have to estimate 1024 likelihoods of feature configurations for each class. Assuming class-conditional independence reduces the number of required likelihoods from 1024 to 8. Another benefit is that class-conditional independence allows inferences about new feature configurations. Even if a particular combination of feature values has not been observed yet, assuming class-conditional independence allows inference of the likelihood of the feature configuration from the marginal likelihoods of the individual feature values, thereby enabling computing the posterior class probabilities.

3.2.2 Robustness

While class-conditional independence may rarely exactly hold in real-world environments, violations of this assumption do not necessarily impair performance. For instance, a widely used classifier in machine learning is the naïve Bayes model, which treats features as class-conditionally independent and computes the posterior class probabilities accordingly. Both simulation studies and analytic results demonstrate the robustness of this model under a variety of conditions (Domingos & Pazzani, 1997). For instance, if the optimality criterion is classification accuracy (error minimization, i.e., a zero-one loss function), then even if the derived posterior probabilities do not exactly correspond to the true

posterior, as long as the correct category receives the highest posterior probability, classification error will be minimized.

3.2.3 Summary

Treating features as class-conditionally independent in a classification task can be helpful, as it simplifies the problem of parameter estimation and violations of classconditional independence do not necessarily entail a loss in classification accuracy. On the other hand, assuming classconditional independence also puts constraints on the types of classification problems that can be solved. For instance, treating features as class-conditionally independent can make it impossible to solve certain classification problems, such as nonlinearly-separable category structures (Domingos & Pazzani, 1997)

From a psychological perspective, however, presuming class-conditional independence might be a plausible default assumption in category learning. If features are (approximately) class-conditionally independent, this facilitates learning and inference substantially. We designed an experiment to investigate whether people initially presume class-conditional independence, and if people change their beliefs and classification behavior when class-conditional independence does not hold in the environment.

4 Experiment

Our goal was to examine whether people use classconditional independence as a default assumption in category learning when the true environmental probabilities are not known yet, that is, early in learning. In order to test this question, we designed a learning environment in which classification decisions would be strongly different if the learner presumes class-conditional feature independence, rather than basing classification decisions solely on the previous instances with the exact same configuration of feature values.

4.1 Method

4.1.1 Participants

Thirty subjects (mean age 23, $SD = 3.3$ years, 70 % females) participated in a computer-based experiment in exchange for 12 Euro.

4.1.2 Task

Participants' task was to learn classify objects with three binary features into one out of two categories. As stimuli we used simulated biological "plankton" specimens differing in three binary features ("eye", "tail", and "claw", shown in the left image in Figure 1). The classes were labelled as "Species A" vs. "Species B". The assignment of the actual physical features and their values to the underlying

probabilities, as well as the class labels, were randomized across participants.

4.1.3 Procedure

We used a trial-by-trial supervised multiple-cue probabilistic category learning paradigm (e.g. Knowlton, Squire, & Gluck, 1994; Meder & Nelson, 2012; Nelson et al., 2010; Rehder & A. B. Hoffman, 2005). After introducing the task and familiarizing subjects with the three features, on each trial a plankton exemplar with a specific feature value combination was randomly drawn according to the true environmental probabilities (see below) and displayed on the screen. After participants made a classification decision, feedback on the true class was given and the next trial started. Learning continued until criterion performance was achieved. Criterion performance was defined as both (1) an overall classification accuracy of 98 % over the last 100 trials, and (2) accurate classification of the last five instances of every individual configuration of features.

4.1.4 Environment

Using optimal experimental design (OED) principles (Nelson, 2005; J. I. Myung & Pitt, 2009) we conducted simulations to find environmental probabilities that best differentiate between a learner that assumes class-conditional independence and a learner that makes predictions based only on previous instances of the same feature configuration. The possible environmental probabilities for our task consisted of the following parameters: (i) the base rate of Species A (determining the Species B base rate), (ii) the likelihoods of each of the eight possible feature value combinations given Species A and (iii) the corresponding values for Species B. The parameter values were obtained via optimization, using genetic algorithms to search for desirable environments which had frequent configurations of features with large absolute discrepancies between the actual posterior probability of Species A, and the posterior probability presumed based on the class-conditional independence assumption. Formally, the genetic algorithm optimized the following fitness function:

$$\sum_{i=1}^{I} [P_{\text{true}}(C = c \mid F = \boldsymbol{f}_i) - P_{\text{cci}}(C = c \mid F = \boldsymbol{f}_i)]^2 \times P(F = \boldsymbol{f}_i)^2 \tag{3}$$

where i indexes all possible feature value combinations and the subscripts true vs. cci indicate the posteriors calculated according to the true vs. class-conditionally independent parameters.

The obtained environment is summarized in Figure 1. The environment contains five out of eight possible feature combinations (henceforth denoted as 111, 000, 100, 010, 001); the remaining three combinations (011, 101, 110) do not occur. The figure illustrates the category base rates, the likelihoods of the feature configurations given the two classes, as well as the marginal likelihoods of the features, which provide the basis for inferring posterior probabilities according to the class-conditional independence assumption. Note that although nothing in the optimization prescribed finding a deterministic environment, in fact the posterior probabilities of Category A are one or zero, for each of the feature configurations that occurs.

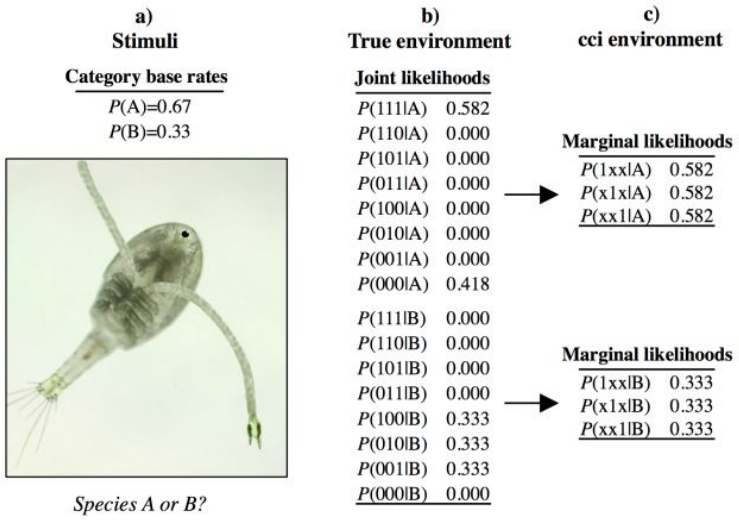

Figure 1. Task environment a) Stimuli, base rates of classes. b) Joint likelihoods of full environment. c) Marginal likelihoods used assuming class-conditional independence

In this environment, assuming class-conditional independence leads to classification decisions that systematically deviate from decisions based on the true environmental probabilities. Table 1 summarizes the feature configurations, their probability of occurrence, the posterior probabilities according to the true environmental probabilities, and the posterior probabilities derived assuming class-conditional independence. For four out of the five feature configurations, the classification decision derived assuming class-conditional independence conflicts with the actual class membership (indicated by ≠).

Table 1

True environment vs. assuming class-conditional independence (cci).

Features			P(features)	P(class | features)			
			true env	true env			with cci
1	1	1	0.39	A	1	=	A 0.91
1	0	0	0.11	B	1	≠	A 0.58
0	1	0	0.11	B	1	≠	A 0.58
0	0	1	0.11	B	1	≠	A 0.58
0	0	0	0.28	A	1	≠	B 0.67

Consider feature configuration 111. This item always belongs to Species A in the true environment. If features are treated as class-conditionally independent, it belongs to Species A with probability 0.91. The small difference between the actual probability of 1.00 and 0.91 should not change the learner's classification decision for this stimulus. This, however, is not true for the other items. For instance, according to the true environment, item 000 belongs to Species A with probability 1, but assuming class-conditional independence entails that it belongs to Species B with probability 0.67. Thus, a learner assuming class-conditional independence would believe that on average about 67 % of the 000 items

belong to Species B, despite experiencing that it always belongs to Species A. The same divergence holds for the other three configurations (100, 010, 001): whereas all of those items actually belong to category B, treating features as class-conditionally independent entails that the probability for category A is higher (0.58).

The strongest discrepancy is for the 000 configuration, which is the second-most-frequent configuration, occurring with probability .28. Note that a hypothetical learner (even with perfect memory) who assumes class-conditional independence of features, and is unable to give up this assumption, will never learn the true statistical structure of this environment, even after completing a quadrillion learning trials. Achieving criterion performance would also be impossible if learners looked at one feature only (at 1xx, or x1x, or xx1 and ignoring the x). Considering single features, participants should think any feature configuration belongs to Species A with probability 0.78. This holds for attending solely to any of the three features.

4.2 Hypotheses

If participants make no (not even tacit) assumptions of classconditional feature independence, and learn each item separately, then items could be learned in order of their frequency of occurrence (a *frequency-of-configuration hypothesis*). If participants approach the task by assuming features to be class-conditionally independent, classification decisions should systematically deviate from ones derived from the true environmental probabilities, especially early in learning (a *posterior-discrepancy hypothesis*).

.Both hypotheses predict the fewest errors for item 111, the most frequent feature configuration and the one for which the class-conditional independence posterior is closest to accurate. For the four critical items, the difference in posterior probability is the largest for item 000. The posterior-discrepancy hypothesis predicts the most errors for item 000, and thus that the ordering of errors should be 111<100≈010≈001<000. However, the frequency-ofconfiguration hypothesis predicts that the ordering of classification errors should be 111<000<100≈010≈001. Key empirical questions are therefore whether there are any systematic differences in learning rate for the individual items, whether the early learning data suggest a presumption of class-conditional independence, and if so, whether the occurrence frequency of an item or the degree to which classconditional independence fails on it determine learning.

4.3 Results and Discussion

All participants reached criterion performance, i.e. learned the category structure (in a mean number of 391 trials, SD=155, Md=348, range 210 to 808 trials). To reach criterion performance, participants needed to classify each individual feature configuration correctly five times in a row. To investigate whether there was a difference in learning speed for the different feature configurations, we calculated the number of times each item needed to be observed before reaching this criterion (Table 2). We will first consider learning time and then error rates.

Table 2

Trials an item needed to be seen to correctly classify it 5 times in a row.

Features			Trials		
			mean	*SD* (*SE*)	median
1	1	1	10.4	10.7 (1.9)	7.0
1	0	0	11.4	8.0 (2.1)	7.5
0	1	0	11.5	7.7 (2.1)	9.0
0	0	1	11.5	7.0 (2.1)	9.0
0	0	0	15.8	11.5 (2.9)	13.5

In our data most subjects learned item 111 before item 000 (22 out of 30, binomial $p < .02$), which is consistent with both hypotheses. Did learning time follow the frequency-of-configuration hypothesis, or the posteriordiscrepancy hypothesis? The posterior-discrepancy hypothesis predicts an ordering of 111<100≈010≈001<000, whereas the item-frequency hypothesis's ordering prediction is 111<000<100≈010≈001. The critical difference in predictions is between the learning time for items 100, 010, and 001 and item 000. The frequency hypothesis predicts that item 000 will be learned faster, whereas the posterior discrepancy hypothesis predicts that items 100, 010, and 001 will be learned first. Here, our results strongly support the posterior discrepancy hypothesis, and contradict the item frequency hypothesis. Items 100, 010 and 001 were learned more quickly by more people than item 000, despite item 000's greater frequency (item 001 faster: 21 out of 30, binomial $p < .05$; item 010 faster: 20 out of 30, binomial $p < .1$; item 100 faster: 21 out of 30, binomial $p < .05$). Moreover, there was a nonsignificant trend for items 100, 010, and 001 to take longer than item 111; consistent with the posterior discrepancy hypothesis but not the configuration frequency hypothesis.

The error rates throughout early learning are summarized in Figure 2. This figure corroborates the analysis of the number of learning trials required for each stimulus configuration: item 000 was clearly the most difficult to learn. As this feature configuration is the one for which the difference in posterior probability is largest when assuming class-conditional independence versus using the full true environmental probabilities, this finding is consistent with the idea that people treat features as being class-conditionally independent early in learning. However, items 100, 010 and 001 were much closer to (or even indistinguishable from) item 111, consistent with the above analysis in Table 2.

5 General Discussion

The present paper examined the role of the assumption of class-conditional independence of features in category learning. While different types of conditional independence assumptions play an important role in various scientific debates and computational models of cognition, little is known about their descriptive validity in the context of classification learning with multiple cues. Our goal was to

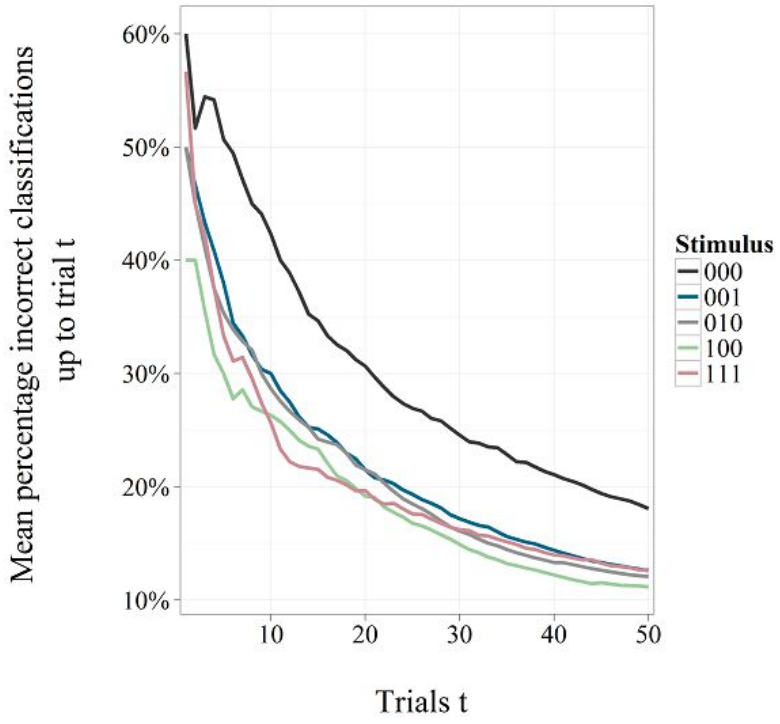

Figure 2. Percentage incorrect classifications for the first 50 trials each item was encountered.

empirically investigate whether people initially (early in learning) treat features as class-conditionally independent. The present results partially support the idea that people initially treat features as class-conditionally independent and make classification decisions accordingly. We think of the results as tentative because some aspects of the data are not perfectly clear.

Our focus in the present study was on participants' behavior early in learning, when evidence about the category structure and environmental probabilities is limited. This approach is similar to the studies of J. D. Smith and Minda (1998), who investigated possible transitions in categorization strategies and stimulus encoding over the course of learning. Their finding was that late in learning exemplar models (e.g., Medin & Schaffer, 1978) accounted best for subjects' behavior, but that this was not the case early in learning (in which a prototype model seemed to better account for human performance, see below). This is also a possible explanation for the finding that despite strongly violating class-conditional independence, the environment in our experiment was clearly learnable. Participants could have initially treated features as class-conditionally independent and computed posteriors accordingly and later shifted to an exemplar-based strategy to minimize classification error.

A key methodological aspect of our study was to use optimal experimental design principles to find environments that would allows us to directly test whether people use class-conditional independence as a default assumption in categorization. Interestingly, the optimizations told us that the best environment to differentiate between a learner that assumes class-conditional independence and a learner that makes predictions based only on previous instances of the same feature configuration was deterministic. The crucial aspect of this environment, however, is not that it is deterministic, but that it entails a nonlinearly separable category structure. Since the class-conditional independence model induces a linear decision bound (Domingos & Pazzani, 1997), it could not achieve criterion performance in this

particular task environment.

This, in turn, relates our study to earlier research in psychology, which investigated whether linearly separable categories are easier to learn than nonlinearly separable ones (Medin & Schaffer, 1978; Medin & Schwanenflugel, 1981). This research focused on two types of categorization models, exemplar- and prototype-models, both of which assume that categorization decisions are derived from similarity comparisons (either to specific exemplars stored in memory or to prototypes of categories). By contrast, we investigated category learning and human subjects' initial assumptions from the perspective of probabilistic inference (see also J. R. Anderson, 1991b; Fried & Holyoak, 1984), a conceptually different view. Nevertheless, there are some interesting connections between our work and these earlier (similarity-based) models. For instance, assuming class-conditional independence entails that not all information (about feature configurations and corresponding class probability) is encoded during learning, but only marginalized conditional likelihoods and category base rates. In this respect the class-conditional independence model is similar to prototype models, which encode parametric information of central tendencies (e.g., mean or mode of feature values) that form the prototype (e.g., J. D. Smith & Minda, 1998).

Importantly, these accounts assume that information is stored separately for each feature and the to-be-classified item is compared to the prototypes separately on each feature dimension individually. Conversely, a learner who makes no assumptions about the structure of the relations between classes and features and directly tracks the true environmental probabilities is conceptually more similar to exemplar models of category learning. The difference is that prototype models, like our independence model, do not need to store each individual instance that is experienced.

In sum, the current paper adds to the debate about the role of conditional independence assumptions for computational models of cognition. The task environment identified based on optimal experimental design principles allowed us to directly examine the descriptive validity of this assumption in category learning. Here, we do find evidence consistent with its use.

6 Acknowledgments

This research was supported by grants NE 1713/1 to JDN and ME 3717/2-1 to BM, from the Deutsche Forschungsgemeinschaft (DFG) as part of the priority program "New Frameworks of Rationality" (SPP 1516). We would like to thank Gregor Caregnato for data collection and Laura Martignon, Michael Waldmann, and Ralf Mayrhofer for their comments.

Jarecki, Meder, & Nelson (submitted)

Naïve and robust: Class-conditional independence in human classification learning

Jana B. Jarecki*, Björn Meder*, Jonathan D. Nelson*

*Max Planck Institute for Human Development, Lentzeallee 94, D-14195 Berlin

Abstract

Research on categorization learning suggests that humans start from relatively simple strategies. This is beneficial in terms of computational efficiency. However, given both the general computational complexity of categorization and the dependence of a strategy's performance on the particular environmental structure, the question arises of when starting simple is beneficial. Drawing on the machine learning literature, we identified one candidate computational principle that human learners could rely on to tackle the computational challenge of probabilistic classification: class-conditional independence of features. This heuristic independence assumption simplifies the inference problem by strongly decreasing the number of parameters required for making probabilistic classification decisions. Also, although in the real world the assumption may not always hold true, acting as if it does often enables accurate and robust classification decisions. Using optimal experimental design principles, we designed and conducted two experiments to investigate whether human category learners initially treat features as class-conditionally independent. We also designed a novel Bayesian classification learning model that infers whether class-conditional independence holds in the environment and adapts its classification decisions accordingly. Most participants' behavior was best accounted for by a model with a high prior belief in class-conditionally independent features. Our findings suggests that class-conditional independence serves as a strong default heuristic assumption in human category learning, which allows learners to make smart and robust inferences.

1 Introduction

Categorization—grouping objects into classes and identifying to which class an object belongs—is a fundamental cognitive ability. From a computational perspective, category learning poses formidable challenges, yet humans excel at it. Cognitive science has investigated how humans induce different category structures, which representations they acquire, and which models best account for learning and generalization. Over the last decades, several categorization models have been suggested, including rule-based (Nosofsky, Palmeri, & McKinley, 1994), prototype (Posner & Keele, 1968), exemplar (Medin & Schaffer, 1978), and probabilistic (J. R. Anderson, 1990; Tenenbaum, 1999) models. Research has also investigated the neural correlates and systems involved in category learning (Ashby, Alfonso-Reese, et al., 1998).

A number of recent studies have focused on dynamic transitions between different categorization strategies during category learning (Medin & E. E. Smith, 1981; J. D. Smith & Minda, 1998; Johansen & Palmeri, 2002; Briscoe & Feldman, 2011; N. D. Goodman et al., 2008; Bourne, Healy, Kole, & Graham, 2006) and effects of sequential presentation (Stewart, Brown, & Chater, 2002). Although there is an ongoing debate regarding which model best describes the early learning phases (e.g., J. D. Smith & Minda, 1998; Johansen & Palmeri, 2002; Bourne et al., 2006), an insight from this line of research is that people start off with rather simple strategies before transitioning to more computationally intense strategies. For instance, Johansen and Palmeri (2002) found that people initially apply unidimensional categorization rules (i.e., make classification decisions based on a single stimulus feature) and adopt more complex rules (i.e., a similarity-based strategy based on multiple features) only if necessary. Similarly, J. D. Smith and Minda (1998) observed that early in learning people tended to use a simple prototype-based strategy and only later shifted to a computationally more demanding exemplar-based strategy (but see Nosofsky & Zaki, 2002, for an alternative interpretation according to which people initially ignore features).

Starting simple is beneficial because it is computationally efficient. A more computationally intense strategy needs to be invoked only if the simple strategy leads to inaccurate class choices. However, whether a particular inference strategy is accurate will vary as a function of the environmental structure (e.g., linearly vs. nonlinearly separable environments; Blair & Homa, 2001; Medin & Schwanenflugel, 1981). How beneficial a rather simple inference strategy can be also depends on whether it performs robustly across different environments.

We approach the problem of robustness and simplicity at the level of the computational challenges that categorization poses. Rather than arguing for or against a particular classification model, we take a probabilistic perspective and investigate the descriptive validity of one particular statistical principle in classification learning: class-conditional independence of features. Treating features as conditionally independent given the true class substantially facilitates the inference problem, while often maintaining excellent classification performance. We examine the hypothesis that class-conditional independence provides a default assumption in the early stages of human classification learning, as it lends support to a simple yet robust inference strategy.

1.1 A computational perspective on classification learning

From a computational standpoint, classification is anything but trivial. One fundamental challenge is the curse of dimensionality (Bellmann, 1961, p. 91): As the number of features and categories in the environment grows, the number of feature–category combinations increases exponentially. Imagine your task is to classify ambiguous, blurry pictures of galaxies[4] into a binary class, namely whether the stars form an elliptical or spiral shape; and assume the galaxies have binary features. Considering galaxies with two, four, or eight features, the number of possible combinations of features and categories grows from $2^3 = 8$ to $2^5 = 32$ to $2^9 = 512$ combinations. The curse of dimensionality is particularly important in real-world situations where the number of features that could be considered is large. Yet, human performance, for example, in classification of image contents where there are many continuous features, has long outperformed computer algorithms (Russakovsky et al., 2015); only recently was an algorithm able to match the best humans (He, Zhang, Ren, & Sun, 2015). Another challenge, especially early in learning, is to make inferences from limited data. Consider, again, galaxy categorization. Your data consist of two samples: one blue and big galaxy, classified as spiral, and one green and small galaxy, classified as elliptical. From this sparse data it is difficult to infer the category of a third, novel object with a different feature configuration (e.g., a blue and small galaxy). In many real-world situations people will not have observed all possible feature–category combinations, necessitating inferences about novel instances from limited information. Yet, humans seem able to readily classify even previously unseen objects, such as galaxies (Lintott et al., 2008).

One way to address the computational challenge of probabilistic classification is the statistical principle of *class-conditional independence*. This principle closely relates to particular cognitive models suggested for early classification behavior (discussed in detail below) and is explicitly represented in the naïve Bayes classifier prominent in machine learning (e.g., Domingos & Pazzani, 1997). Class-conditional independence entails that the probabilities of features are statistically independent given the true class of the object. For instance, if a galaxy belongs to the class of spiral galaxies, learning that it is blue does not provide any information about its other features, such as size.

Our research question is motivated by results from machine learning showing that it can be efficient in many situations to assume class-conditional feature independence. First, doing so diminishes the curse of dimensionality by simplifying the required inference. Second, even when features are *not* conditionally independent given the true class, treating features *as if* they are does not necessarily impair classification performance (e.g., Domingos & Pazzani, 1997). Thus, treating features as class-conditionally independent can enable robust inferences even when there is a mismatch with the true statistical structure of the environment.

1.2 Scope and goal of the present study

The goal of the present work was to investigate whether class-conditional independence serves as a heuristic default assumption in human category learning. We investigated whether humans sidestep the aforementioned computational challenges with the default assumption that features are class-

[4] A task that was actually done by more than 150,000 people on www.galaxyzoo.org.

conditionally independent. To this end, we implemented the assumption in a probabilistic model of category learning that enabled us to determine to what extent individual people behave as if they make this assumption, and how people's inferences dynamically adapt to the statistical structure of the environment.

The present paper complements existing research in a number of ways. Our work allies with recent evidence demonstrating interindividual differences in category learning (McDaniel et al., 2014; Bartlema, Lee, Wetzels, & Vanpaemel, 2014). We modeled individual learning trajectories and beliefs about environmental structure. To specifically investigate variation in individuals' learning speed, we used a learning paradigm that does not terminate after a fixed number of trials (e.g., Nosofsky, Kruschke, & Mckinley, 1992; Nosofsky, Palmeri, & McKinley, 1994; J. D. Smith & Minda, 2002; Erickson & Kruschke, 1998) or combinations of trials and performance (Medin & Schwanenflugel, 1981; Mckinley & Nosofsky, 1995); our study used only a performance criterion (like e.g., Homa, Dunbar, & Nohre, 1991; Medin & E. E. Smith, 1981). One limitation of previous studies arguing that people shift from simple to more complex classification strategies is that they focused on rather large, discrete bins of learning trials (56 in J. D. Smith and Minda, 1998; 36 in Johansen and Palmeri, 2002). We sought to dynamically model the learning phase for each trial. Further, rather than inserting test trials in the learning phase to track the learning process (Nosofsky, Kruschke, & Mckinley, 1992; Nosofsky, Palmeri, & McKinley, 1994; J. D. Smith & Minda, 2002; Erickson & Kruschke, 1998), we designed our task such that it optimally differentiated our hypotheses and allowed us to present all exemplars during learning. We used optimal experimental design principles (J. I. Myung & Pitt, 2009; Nelson, 2005; Nelson et al., 2010) to design a task in which learners who presume class-conditional independence will arrive at strongly different classification decisions from those made by fully flexible learners who do not make this assumption. Hence, the observed decisions during learning provide insights regarding whether humans assume that features are independent given the class, prior to experiencing the task.

1.3 Class-conditional independence

Generally, from a probabilistic modeling perspective, classification requires estimating the probability that a given stimulus s_j belongs to the i^{th} class c_i, which can be computed using Bayes's rule:

$$p(C = c_i | S = s_j) = \frac{p(S = s_j \mid C = c_i) p(C = c_i)}{p(S = s_j)}, \tag{4}$$

where $C \in \{c_1, \ldots, c_n\}$ denotes the class random variable, and $S \in \{s_1, \ldots, s_m\}$ the stimulus random variable. Each stimulus s_j represents one possible feature configuration. In the following, we omit capital letters and most subscripts and denote the probability of a class given a feature configuration as

$$p(c \mid s) = \frac{p(s \mid c)\, p(c)}{p(s)}, \tag{5}$$

where $p(s \mid c)$ denotes the stimulus likelihood (the likelihood of a feature configuration) given the class, and $p(c)$ denotes the class base rate (probability of the class). The denominator $p(s)$ is a normalizing constant given by $\sum_i p(s \mid c_i) * p(c_i)$.

Class-conditional feature independence means statistical independence of the features given the true class (e.g., Rish et al., 2001; Domingos & Pazzani, 1997; Flach & Lachiche, 2004). Two random variables are statistically independent if knowing the value of one variable does not provide information about the other variable, or in other words, if their joint probability equals the product of their marginal probabilities. This can be used to compute the stimulus likelihoods $p(s \mid c)$ in Equation 5:

$$p(s \mid c) = \prod_{d=1}^{D} p(f_d \mid c), \tag{6}$$

where s denotes the stimulus, c the class, f_d the d^{th} feature of this stimulus, and D the total number of features. For example, when classifying coffees as expensive or cheap by origin and roasting method, we have $D = 2$ features.

Let us use the coffee example for illustration: According to Equation 6, the stimulus likelihood of a light-roasted coffee from Brazil (given it belonged to the class 'expensive') can be inferred by multiplying the likelihood of any light-roasted coffee—independent of its origin—with the likelihood of any coffee from Brazil—independent of its roast (given it was expensive). The probability $p(f_d \mid c)$ is called the *marginal feature likelihood* of feature number d given class c. Let us illustrate the difference between *marginal* likelihoods and *configural* likelihoods by, again, using the example of classifying coffees as expensive or cheap from the two binary features origin and roasting method. The probability that a coffee originated from Brazil, independent of roast, given it was cheap, $p(Brazil \mid cheap)$ constitutes a *marginal feature* likelihood. The probability that a coffee originated from Brazil *and* had a light roast, given it was cheap, $p(Brazil \;\&\; light\ roast \mid cheap)$, constitutes a *configural stimulus* likelihood. Thus, a configural stimulus likelihood refers to a combination of several features given a class, whereas the marginal feature likelihood refers to the likelihood of a single feature given the class, marginalizing over the other features.

1.4 Class-conditional independence in cognitive science

As detailed above, the exponential growth of feature–category combinations poses a challenge. How do classification models address this, and how do they relate to class-conditional independence? Some probabilistic models state conditional independence explicitly (e.g., Barrington, Marks, Hsiao, & Cottrell, 2008; D. Friedman, Massaro, Kitzis, & Cohen, 1995; J. R. Anderson, 1991a; Shafto, Kemp, Mansinghka, & Tenenbaum, 2011), whereas it is implicit in other models. Consider decision-boundary models, where linear decision bounds combine cues in a weighted-additive manner and ignore interactions, thereby addressing the curse of dimensionality. Linear feature separability is related to class-conditional feature independence, because in log space the naïve Bayes classifier is an interaction-

free additive model (Zhang & Ling, 2001; Manning, Raghavan, & Schutze, 2009). A naïve Bayes classifier is restricted to linearly separable features in the case of binary features, but not necessarily for features with more than two values (Zhang & Ling, 2001). Fast-and-frugal tree models (Martignon, Katsikopoulos, & Woike, 2008; Luan et al., 2011) use a pruned-tree structure in which they consider one feature at a time, thereby ignoring feature interactions.[5] Finally, additive prototype models (e.g., Reed, 1972; Posner & Keele, 1968) compare the current stimulus to the most typical stimulus in each category. Since there is only one prototype per category, the number of comparisons between a novel stimulus and the prototypes increases only linearly in the number of categories (instead of exponentially in the number of features). Note that—besides fast-and-frugal trees, which have not been so widely studied—the models addressing the curse of dimensionality are also the models hypothesized to describe early human learning (J. D. Smith & Minda, 1998; Johansen & Palmeri, 2002).

Importantly, in this line of research, the assumption that people assume a particular feature-dependency structure at the beginning of learning was at most indirectly addressed. To test this more directly, we need to pit a model that initially explicitly assumes class-conditional independence against one without this prior assumption and allow the model to learn the dependency structure by experience.

Conditional independence has been investigated more explicitly in research on causal learning and reasoning. A central assumption of causal Bayes nets theory (Pearl, 2000; Spirtes et al., 1993; Sloman, 2005) is the causal Markov condition (Spirtes et al., 1993), which states that a variable in a causal network is independent of all other variables, conditional on its direct causes, except its causal descendants.[6] The empirical evidence as to whether people honor the Markov condition is mixed. While some studies support key normative principles underlying causal Bayes nets theory (e.g., Meder, Hagmayer, & Waldmann, 2008; Meder, Hagmayer, & Waldmann, 2009; Meder, Mayrhofer, & Waldmann, 2014), other studies have found violations of conditional independence in causal reasoning (Rehder & A. B. Hoffman, 2005; Rehder, 2014), which may reflect participants' causal knowledge (Walsh & Sloman, 2008) or result from contextual information brought to the task (Mayrhofer & Waldmann, 2014). Thus, from the causal reasoning literature, it is an open question whether and under which conditions people honor class-conditional independence.

1.5 Benefits of assuming class-conditional independence

Assuming that features are class-conditionally independent has computational benefits. The key advantage of making this assumption is that it addresses the curse of dimensionality. Specifically, assuming class-conditional independence reduces the number of parameters a probabilistic model needs to estimate (for Equation 5), compared to assuming flexible feature dependencies. These parameters are the stimulus likelihoods $p(s \mid c)$ and the class base rates $p(c)$. They can be estimated from the

[5] Of course, the nodes in the tree could also consist of feature configurations, rather than individual features. In this case, however, the problem of the many possible feature configurations again applies.

[6] The close relationship between class-conditional independence and the Markov condition is best illustrated with a common-cause network. Consider a binary cause C with three binary effects, E_1, E_2, and E_3. Applying the causal Markov condition to this causal structure entails that the three effects are independent of each other conditional on their common cause C. Now, if C represents a binary class variable and E_1, E_2, and E_3 represent three binary features, the assumption of class-conditional independence is equivalent to the causal Markov condition. Thus, class-conditional independence can be considered a special case of the Markov condition applied to a common-cause model.

data (experienced stimuli and feedback about the true class), but if the data is scarce and noisy and the number of to-be-estimated quantities is large, the resulting estimates tend to be imprecise or may not be available. Inferring the class base rate $p(c)$ is relatively straightforward, because usually the class has only few values and the relevant information is obtained each time the learner receives class feedback. By contrast, inferring the stimulus likelihoods $p(s \mid c)$ becomes increasingly difficult with more features.[7] The number of stimuli (feature configurations) in a task with D binary features is 2^D, therefore growing exponentially with the features. This complicates estimating the stimulus likelihood. More generally, the total number of parameters a probabilistic model requires for a binary classification with D features is $2^{D+1}-1$, if the model allows for interactions among features given the class. Assuming class-conditional independence reduces this exponential growth to a linear growth; in this case the total number of necessary parameters is only $2D+1$. Figure 3 illustrates this difference in parameter growth. Thus, treating features as class-conditionally independent reduces the computational complexity by reducing the number of quantities required for Equation 5.[8]

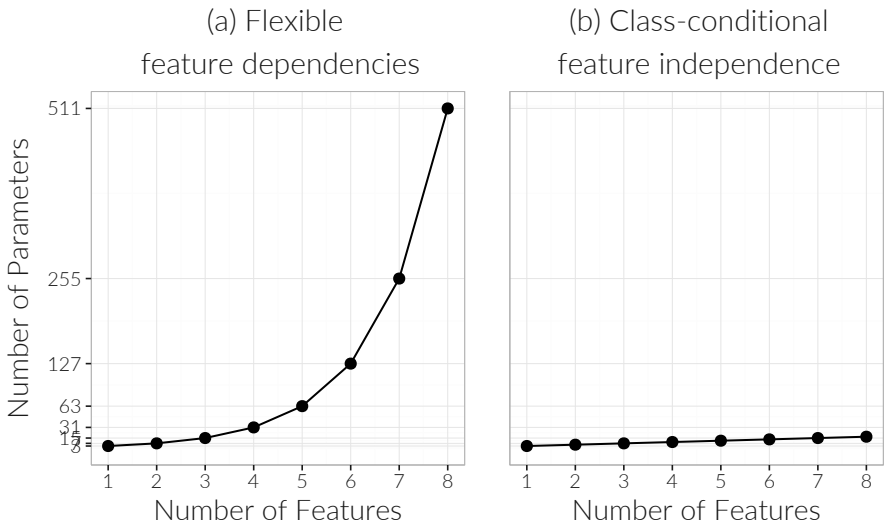

Figure 3. Growth of the number of parameters with features. Parameter growth in a binary categorization task as the number of features increases: **(a)** probabilistic model allowing for flexible feature dependencies; **(b)** probabilistic model relying on class-conditional feature independence.

A second benefit of the presumption of class-conditional independence is the ability to make inferences about new feature configurations, beyond what can be inferred by the category base rates alone. Even if a particular stimulus (i.e., a new configuration of feature values) has not been observed yet, the individual features that constitute it may have been observed before. Returning to the coffee example, if two coffees with the features "$Brazil$ & $light\ roast$" and "$Columbia$ & $dark\ roast$," and the associated class, say 'expensive', have been observed, class-conditional independence enables

[7] To illustrate, consider a stimulus with binary feature values. How many possible stimuli (feature configurations) exist if there are two, three, or four features? Two features yield $2^2 = 4$ stimuli, three features yield $2^3 = 8$ stimuli, four features yield $2^4 = 16$ stimuli, and so forth.

[8] Generally (beyond binary classes and features), for a class vector $\boldsymbol{c}$ and D features, the number of parameters is $(\prod_{d=1}^{D}|\boldsymbol{f}_d| - 1)\cdot|\boldsymbol{c}|$, where $|\boldsymbol{c}|$ denotes the number different classes, D is the number of different features, and $|\boldsymbol{f}_d|$ the number of the values the d^{th} feature can take. By contrast, if class-conditional independence holds, the number of parameters is $\sum_{d=1}^{D}(|\boldsymbol{f}_d|-1)\cdot|\boldsymbol{c}|$.

inferences about the unseen feature configuration "$Brazil$ & $dark\ roast$" by using the marginal feature likelihoods of the experienced coffees to compute the configural likelihood of the new coffee according to Equation 6.

Another advantage of class-conditional independence is its robustness. Although class-conditional independence may not hold exactly in many real-world environments (Titterington et al., 1981; Rish et al., 2001), relying on the assumption and making classification decisions accordingly does not necessarily impair performance. Both simulation studies and analytic results demonstrate the robustness of the naïve Bayes classifier, which treats features as class-conditionally independent and often arrives at accurate class decisions even in tasks where features are strongly correlated given the class (Domingos and Pazzani, 1997; for other conditions under which the classifier performs well, see Rish et al., 2001). One reason for the robustness of the model is that in many cases it assigns the highest probability to the most probable class (i.e., uses an arg max decision rule), even if it misjudges the exact numerical probability estimate (see e.g., Domingos & Pazzani, 1997; Flach & Lachiche, 2004). The classifier will, for example, classify a coffee as "expensive" if the probability for this category is $p = .60$ and also if it is $p = .70$. As long as the computed class probability exceeds the binary decision threshold of .50, the class decision remains the same. Another reason for the good performance of the naïve Bayes model is that it is less prone to overfitting. More complex models are more likely to fit noise in the data and therefore tend to generalize poorly when predicting from little training data. Simpler models may be associated with higher bias, but they tend to be less sensitive to noise in the data. The trade-off between bias and variance is known as the bias–variance dilemma of prediction (Geman, Bienenstock, & Doursat, 1992) and has been generalized to classification errors (J. H. Friedman, 1997). It has been argued that exemplar models tend to be biased, whereas prototype models tend to overfit complexity relative to humans (Briscoe & Feldman, 2011). Like the former, the naïve Bayes algorithm is associated with high bias and low variance (J. H. Friedman, 1997).

In sum, assuming that features are independent given the class has several computational benefits. We therefore hypothesized that class-conditional independence may be a guiding principle in human classification learning, too. From a psychological perspective, treating the environment as class-conditionally independent is a "bet" that the task structure complies with it. Acting according to this bet might be useful, because it reduces cognitive effort and is robust against violations of the assumption. However, dogmatically adhering to class-conditional independence in the face of substantial contradictory evidence would clearly be a bad idea, since it severely constrains the type of category structures that can be learned. For instance, it prevents learning the exclusive-OR problems that have been extensively studied in machine learning (Minsky & Papert, 1969), human categorization (Love et al., 2004; Little & Lewandowsky, 2009), and causal learning (Waldmann & Martignon, 1998; Walsh & Sloman, 2008).[9] Humans can learn exclusive-OR problems (e.g., Little & Lewandowsky, 2009) and with sufficient experience can learn inferences based on configural stimuli rather than marginal features (e.g., Nosofsky & Bergert, 2007; Little & Lewandowsky, 2009; Johansen & Palmeri, 2002).

[9]For instance, if coffee is "expensive" if it is either from Brazil or lightly roasted, but not when it is from Brazil *and* lightly roasted and also not when it is neither from Brazil nor lightly roasted, this class structure is in line with exclusive-OR. More formally, if an object belongs to class $C = 1$ if it has either feature $f_1 = 1$ or feature $f_2 = 1$, but not when both or neither of the two features is present.

Therefore, our hypothesis is not that human category learners always assume class-conditional independence, no matter what they have experienced. Rather, the idea is that learners use this principle as a default assumption when facing novel classification tasks.

In the following, we first introduce the two statistical task environments used to address our research questions. Then we present a novel Bayesian model that takes into account the uncertainty of the classification environment, that is, uncertainty regarding whether class-conditional independence holds. We then present our study of classification learning, both by our model learners and by human subjects, in both of these environments. We conclude by discussing the implications of our findings for models of category learning and probabilistic inference.

2 Design: Statistical Task Environment

We used a classification task involving three binary features. The statistical task structure was designed using optimal experimental design principles (J. I. Myung & Pitt, 2009; Nelson, 2005). Our goal was to identify a statistical environment that best differentiates between learners who do and do not assume class-conditional independence, such that a probabilistic model that treats features as class-conditionally independent makes maximally diverging decisions from a model that does not treat features as such.

2.1 Identifying task environments

In a task with three binary features and a binary class, the statistical environment is defined by the base rate of class 1, $p(c_1)$; the likelihoods of the eight stimuli (feature configurations) given class 1, $p(s_1 \mid c_1), \ldots, p(s_8 \mid c_1)$; and the corresponding likelihoods given class 2. The class 2 base rate is $1 - p(c_1)$. Given these quantities, the posterior probability with which each stimulus belongs to class 1, or to class 2, can be computed via Equation 5. We can then derive the marginal feature likelihoods and recompute the probability that the stimuli belong to class 1 using class-conditional independence (Equation 6). We employed numeric optimization methods using a genetic algorithm to find environmental probabilities that implied different class decisions depending on whether the posterior probabilities were computed presuming class-conditional independence or not (see A.2.1 for details).

2.2 Environment 1: Deterministic task

Table 3 shows the resulting task environment. It contains five stimuli, which we henceforth denote as 000, 001, 010, 100, 111; three of the eight possible stimuli (110, 101, 011) do not occur. Note that this was a result of the optimization, not a deliberate choice on our part. A linear combination of the features cannot predict the correct class, which is illustrated in Figure 4, in which no plane can separate the classes perfectly. A classifier assuming class-conditional independence (i.e., the naïve Bayes model) cannot learn this task structure; nor can a linear classification rule without interaction terms (Ashby & Townsend, 1986); nor a pure prototype model (Reed, 1972); nor a fast-and-frugal tree (Martignon,

Vitouch, Takezawa, & Forster, 2003). This task structure allows us to test whether people assume class-conditional independence, because it predicts strongly divergent classification decisions depending on whether class-conditional independence is assumed.

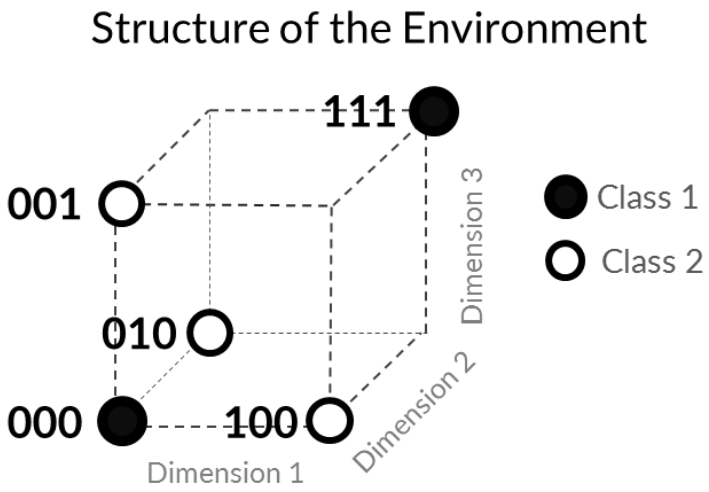

Figure 4. Schematic representation of the environmental structure. The axes denote the stimulus dimensions of the three-dimensional space, dots represent feature configurations (stimuli), and numbers denote binary feature values on the dimensions (e.g., stimulus 100 has value 1 on the first dimension and value 0 on the second and third dimension).

Table 3 summarizes Environment 1. It shows the stimuli, how frequently they occur, the true probability with which each stimulus belongs to class 1, and the corresponding probability derived assuming that the features are independent given the class. For four of the five stimuli the classification decision based on class-conditional independence conflicts with the actual class membership (indicated by $\neq$). We refer to those items as *critical stimuli*. Note that this task is deterministic, as the probability of belonging to either of the two classes is 1 or 0. Figure 5, below, summarizes all parameters that describe the task (the class base rate, the configural stimulus likelihoods, and the corresponding marginal feature likelihoods).

The task environment entails that a probabilistic model that assumes class-conditional independence selects a different class from that chosen by a probabilistic model that knows the true stimulus likelihoods. Only for one of the five stimuli (feature configuration 111) does the same classification decision result. Stimulus 111 belongs to class 1 with probability 1 in the actual environment; presuming class conditional independence, it belongs to class 1 with probability .91. The difference of 9 percentage points does not change the fact that both models select class 1, which is the most likely class.

This equivalence in model decisions, however, does not hold for the remaining four critical stimuli. In the actual environment, stimulus 000 belongs to class 1 with probability 1, but under class-conditional independence it would belong to class 2 with probability .67. Similarly, the other critical stimuli (100, 010, 001) actually belong to class 2, but a learner assuming class-conditional independence would assign them to class 1, because $p(c_1 \mid s) = .58$. The model disagreement is strongest for stimulus 000, with

Table 3

Environment 1 (deterministic task)

Stimulus s			$p(s)$	$p(c_1 \mid s)$		
			true env	true env		with CCI
1	1	1	.39	1	$\approx$	.91
1	0	0	.11	0	$\neq$	.58
0	1	0	.11	0	$\neq$	.58
0	0	1	.11	0	$\neq$	.58
0	0	0	.28	1	$\neq$	.33

Note: Shown are the stimuli s, their probability of occurrence $p(s)$, the true posterior class probabilities, and the posterior class probabilities derived assuming class-conditional independence (CCI). In this task, assuming class-conditional independence yields classification decisions that deviate from those based on the true task parameters (rows indicated by $\neq$), because the true parameters involve more flexible feature dependencies. env = environment.

$p(c_1 \mid s; true\ probabilities) = 1.00$ but $p(c_1 \mid s;\ cci) = .33$. Note that the stimuli are not equally frequent. The uncritical stimulus 111 is most frequent and stimulus 000 is the second-most-frequent configuration.

A model that assumes features are conditionally independent will perform poorly in this environment, no matter how many learning trials it encounters. The poor performance is not a result of a lack of training data, rather, it follows from (falsely) assuming that features are class-conditionally independent, a structural assumption that cannot be changed by more learning input. A learner (even with perfect memory) who believes in class-conditional independence will not learn the true statistical structure of the environment as long as the learner is unable to give up this belief, even after experiencing infinite numbers of training examples.

An interesting property of this environment is that class membership is deterministic, as for all stimuli $p(c \mid s) = 1$ in the actual environment. Although our optimization did not explicitly aim for this, from a mathematical perspective the best way to differentiate whether a classifier presumes class-conditional independence is to use a deterministic classification task. However, we do not want to limit our analyses and empirical results to such environments, which we suspect are fairly rare in the real world. Also, research comparing learning in deterministic and probabilistic tasks (e.g., Mehta & Williams, 2002; Little & Lewandowsky, 2009) found that participants needed longer to learn probabilistic category structures (but see Seger & Cincotta, 2005). We therefore designed a second, probabilistic environment. In this environment, too, treating features as class-conditionally independent leads to different behavior versus directly learning the configural feature likelihoods.

2.3 Environment 2: Probabilistic task

We manually changed the parameters of the first environment to design a probabilistic analogue of Environment 1. This second environment involves the same five stimuli and frequencies as the first environment (except for a 1 percentage point difference for stimuli 111 and 000). The precise numerical probabilities with which the stimuli belong to the classes differ from the first environment, but the divergences between classification decision when assuming class-conditional independence versus encoding the true stimulus likelihoods are identical. Stimulus 000 belongs to class 1 in the true environment but would be assigned to class 2 when assuming class-conditional independence. Stimuli 001, 010, and 100 actually belong to class 2, whereas a learner assuming class-conditional independence would assign them to class 1. Table 4 shows the second, probabilistic environment.

Table 4

Environment 2 (probabilistic task)

Stimulus s			$p(s)$	$p(c_1 \mid s)$		
			true env	true env		with CCI
1	1	1	.38	.95	$\approx$	.89
1	0	0	.11	.25	$\neq$	.65
0	1	0	.11	.25	$\neq$	.65
0	0	1	.11	.25	$\neq$	.65
0	0	0	.29	.94	$\neq$	.48

Note: Here, the to-be-learned class probabilities $p(c_1 \mid s)$ are probabilistic instead of deterministic. Shown are the stimuli, their frequency, the true posterior class probabilities, and the posterior class probabilities derived assuming class-conditional independence (CCI). Assuming class-conditional independence yields classification decisions deviating from those based on the true task parameters (rows indicated by $\neq$). env = environment.

Figure 5 displays a comparison of the parameters required to specify the task environment depending on whether class-conditional independence is assumed. The comparison shows that the number of parameters is smaller when assuming class-conditional independence: Only three marginal feature likelihoods for each class are required; in total, seven probabilities need to be estimated from data. Without the assumption of class-conditional independence, seven configural stimulus likelihoods for each class need to be estimated (the eighth stimulus likelihood is implied because the likelihoods given one class sum to 1); in total, 15 probability estimates are required.

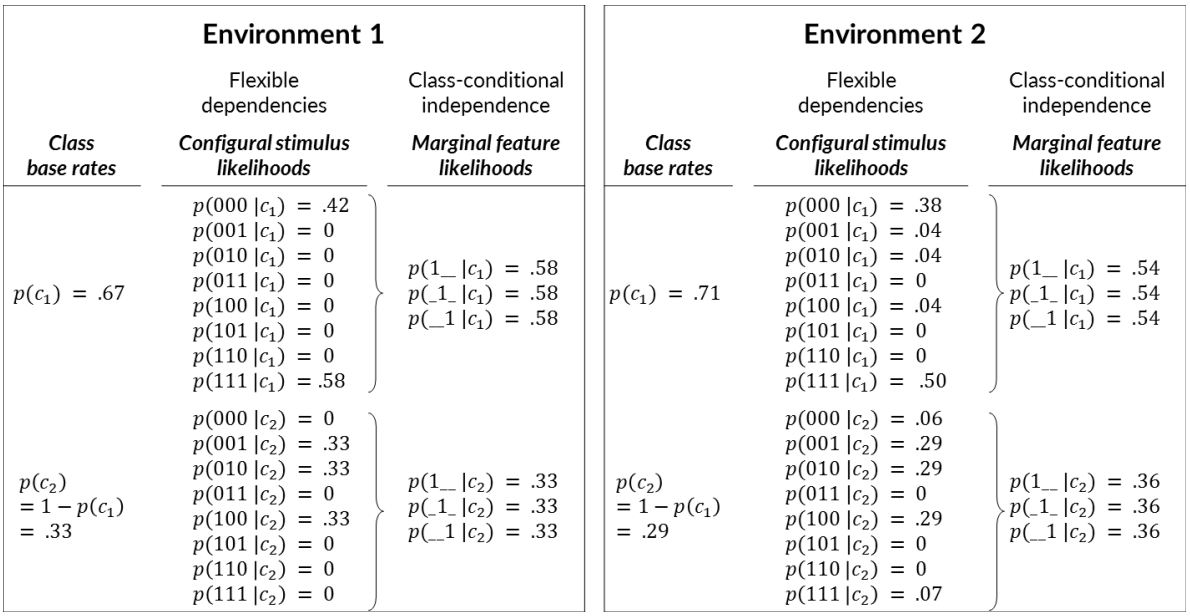

Figure 5. Complexity reduction with class-conditional independence. Assuming class-conditional feature independence reduces the number of class-conditional stimulus probabilities that are needed to describe the environment. The figure shows two ways to describe the classification task, with and without the assumption of class-conditional independence. The *class base rate* is required in either case. The Flexible dependencies column shows that eight likelihoods (or class-conditional stimulus probabilities) describe the environments. The Class-conditional independence column shows that only three marginal likelihoods are required under the assumption of class-conditional independence. The environment in the left panel is deterministic, the one in the right panel probabilistic. *Note.* $p(c_1)$ = class 1 base rate, $p(000 \mid c_1)$ = probability of configural stimulus 000 given class 1, $p(1 \mid c_1)$ = probability of first marginal stimulus dimension given class 1.

In our experiments and simulations we embedded the statistical task environments in a trial-by-trial learning task. Models and human learners were presented with one stimulus, randomly drawn from the task distribution, and they received feedback about the true class after their classification decision. For about two thirds of all decisions different classifications would result, depending on the assumption of class-conditional independence (in the limit, assuming knowledge of the relevant probability estimates), in both task environments.

3 The probabilistic dependence/independence structure and category-learning model (DISC-LM)

We next describe the probabilistic dependence/independence structure and category-learning model (DISC-LM). The model incorporates the uncertainty about whether features are independent given the class, and the uncertainty about the feature likelihoods and the class base rate. The model computes the probability that the current stimulus belongs to class 1 twice: On the one hand, it computes the

stimulus likelihoods according to class-conditional feature independence, $p(c \mid s; cci)$, and on the other hand according to flexible conditional feature dependencies, $p(c \mid s; flex)$. Remember that the resulting class decisions need not agree. The model weights the obtained posterior probabilities according to the match between the data and the structural assumption about feature independence Bayesian model averaging, Chickering and Heckerman, 1997. The resulting classification reflects both the uncertainty about whether features are class-conditionally independent and the uncertainty within each structural model about the true values of the class base rate and the stimulus likelihoods. By relating the model behavior to empirical data it is possible to investigate whether human classifiers treat features as class-conditionally independent.

The DISC-LM infers the class base rate and the stimulus likelihoods in a Bayesian way. We employed Monte Carlo simulations to numerically estimate the required probability densities. The simulations and analyses were programmed in R (R Core Team, 2014). (For a list with symbols and notation, see A.2.2. A.2.3 and A.2.4 detail the steps of the Monte Carlo simulation process.)

3.1 Model parameters

The DISC-LM has two parameters. The *structural belief* parameter π governs the model's prior belief that the structure of the environment complies with class-conditional independence. Higher values of π lead to a stronger influence of the posterior class probability derived under the assumption of class-conditional independence when integrating out the two estimates. Importantly, different values of the structural belief parameter can be used to model individual differences in believing that features are class-conditionally independent. The *conservatism* parameter δ governs how much experience the model requires to learn the probabilities needed for computing the class probabilities. Higher values of δ lead to more conservative learning. The conservatism parameter enables DISC-LM to account for interindividual differences in learning speed.

3.1.1 Inferring the class base rate

The model infers the class base rate $p(c_1)$ separately from the stimulus likelihoods. Using standard Bayesian inference (see e.g., Griffiths, Kemp, & Tenenbaum, 2008), the model integrates prior knowledge about the class with the information about the class gained from experience with the task. The former is the prior probability $p(c_1)$. Let $E_t = e_1, \cdots, e_t$ denote all classes experienced until trial t. Let capital P denote densities. Given E_t, the model integrates the density of the prior probability $P(p(c_1))$ with the density of the likelihood $P(E_t \mid p(c_1))$, and forms a posterior density over possible class base rates given the experience, $P(p(c_1) \mid E_t)$. We refer to the resulting distribution as the posterior distribution of the class base rate, given by

$$P(p(c_1) \mid E_t) = \frac{P(E_t \mid p(c_1)) \cdot P(p(c_1))}{P(E_t)}, \tag{7}$$

where $P(E_t) = P(E_t \mid p(c_1))\, P(p(c_1)) + P(E_t \mid p(1-c_1))\, P(1-p(c_1))$.

The model's prior distribution was set to a uniform Beta distribution.

$$P(p(c_1)) = Beta(\delta, \delta) \tag{8}$$

where δ is a conservatism parameter. With higher values of δ the posterior distribution of $p(c_1)$ has a higher density for values around .50.[10] In other words, the larger δ is, the more evidence is required to shift the posterior distribution toward the observed distribution of the two classes (i.e., to the maximum-likelihood estimate of $p(c_1)$).

3.1.2 Inferring the stimulus likelihoods with flexible feature dependencies

The DISC-LM infers the posterior class probability in two ways, which differ in how the model learns the stimulus likelihoods $p(s \mid c_1)$ and $p(s \mid c_2)$. The first way is according to *flexible class-conditional feature dependencies*, meaning that the model directly encodes the likelihoods of the eight stimuli. Let us denote both likelihoods by the shorthand $p(s \mid c)$. Let $E_t = e_1, \ldots, e_t$ denote the stimuli (feature configurations) given the class that the model has experienced until trial t. Given E_t, the model infers the likelihoods of all eight stimuli $p(s \mid c) = p(s_1 \mid c), \ldots, p(s_8 \mid c)$ by integrating the prior belief about the stimulus likelihoods $P(p(s \mid c))$ with how likely it is that the experienced data were generated by a particular stimulus likelihood $P(E_t \mid p(s \mid c))$. This yields a distribution of the posterior belief about the stimulus likelihoods $P(p(s \mid c) \mid E_t)$:

$$P(p(s \mid c) \mid E_t) = \frac{P(E_t \mid p(s \mid c)) \cdot P(p(s \mid c))}{P(E_t)}, \tag{9}$$

where $P(E_t) = \sum_{i=1}^{2} P(E_t \mid p(s \mid c_i)) \cdot P(p(s \mid c_i))$.

The prior distribution of the stimulus likelihoods, $P(p(s \mid c))$, was set to a uniform Dirichlet distribution:

$$P(p(s \mid c)) = Dirichlet(\delta, \delta, \delta, \delta, \delta, \delta, \delta, \delta). \tag{10}$$

We have two separate distributions, one for inferring the likelihood given c_1, another given c_2. As before, the conservatism parameter δ governs the speed of learning. Higher values of δ imply a posterior distribution that favors the hypothesis that all stimuli are equally likely, $p(s_1 \mid c) = p(s_2 \mid c) = \ldots = p(s_8 \mid c) = \frac{1}{8}$. The parameter values of δ were equal for base-rate and likelihood estimates.

3.1.3 Inferring the stimulus likelihoods with class-conditional feature independence

The second way in which DISC-LM infers the stimulus likelihoods is according to class-conditional feature independence. In this case the model learns the marginal feature likelihoods, $p(f_d \mid c)$, for each

[10] To illustrate, compare inferences given $\delta = 1$ or $\delta = 100$. Suppose we have observed class 1 in one trial. After this observation, the mean of the posterior density of the class 1 base rate given $\delta = 1$ equals $\frac{2}{2+1} = 0.6667$, whereas it equals $\frac{101}{101+100} = 0.5025$ given $\delta = 100$.

of the $d = 1,2,3$ features and multiplies them to arrive at the configural stimulus likelihoods according to Equation 6. Let $E_{td} = e_{1d},\ldots,e_{td}$ denote the values of the d^{th} feature given class c that were experienced up until trial t. Given this data, the model integrates its prior belief about the likelihood of feature d, $P(p(f_d \mid c))$, with the likelihood of the experience given all likelihoods of this feature, $P(E_t \mid p(f_d \mid c))$, to form the posterior distribution of the likelihoods of feature d, $P(p(f_d \mid c) \mid E_t)$:

$$P(p(f_d \mid c) \mid E_t) = \frac{P(E_t \mid p(f_d \mid c)) \cdot P(p(f_d \mid c))}{P(E_t)}, \tag{11}$$

where $P(E_t) = \sum_{i=1}^{2} P(E_t \mid p(f_d \mid c_i)) \cdot P(p(f_d \mid c_i))$. The prior distributions of the likelihoods of the individual features d were set to independent uniform Beta distributions:

$$P(p(f_d \mid c)) = Beta(\delta, \delta) \qquad \forall\, d = 1,2,3, \tag{12}$$

where δ is the conservatism parameter. The higher δ is, the closer the posterior belief about the feature likelihood will be to .50.

The posterior densities of the feature likelihoods are then multiplied, yielding a posterior density of the configural stimulus likelihoods, $P(p(s \mid c) \mid E_t)$. The stimulus likelihoods converge toward the true likelihoods if class-conditional independence holds in the environment. If this is not the case, the inferred configural likelihoods can deviate from the true stimulus likelihoods.

The predictions derived from flexible conditional feature dependencies and class-conditional feature independence are then combined using Bayesian model averaging (see also A.2.5). The model's final estimate of the probability that stimulus s belongs to class c, is given by a weighted average:

$$\hat{p}(c \mid s) = w\, \hat{p}(c \mid s; cci) + (1 - w)\, \hat{p}(c \mid s; flex), \tag{13}$$

where $flex$ and cci denote whether the point estimates were generated assuming flexible feature dependencies or class-conditional feature independence, respectively, and w is the posterior structural belief that weights the point estimates, $0 \leq w \leq 1$.

The posterior structural belief w captures how strongly the model believes that class-conditional independence holds. w is derived by combining the prior structural belief about whether class-conditional independence holds, π, with the likelihood of the data given class-conditional independence. Let $E_t = e_1,\ldots,e_t$ be the experienced data (joint values of stimuli and classes) until trial t. Given E_t, the model integrates the prior structural belief with the likelihood to form the posterior structural belief $w(t)$ in each trial t as follows:

$$w(t) = \frac{P(E_t \mid p(s,c;cci)) \cdot \pi}{P(E_t)}, \tag{14}$$

where π denotes the *prior* structural belief, $0 \leq \pi \leq 1$, and is a probability instead of a distribution.

The term $p(s,c;cci)$ is the joint prior probability of stimuli and classes if the stimulus likelihoods are inferred using class-conditional independence (Equation 11). The denominator $P(E_t)$ is a normalizing constant given by $P(E_t) = P(E_t \mid p(s,c;cci))\pi + P(E_t \mid p(s,c;flex))(1-\pi)$, where $flex$ denotes that likelihoods were inferred using no independence assumption (Equation 9).

The likelihood that the experienced data were generated by features that are class-conditionally independent is given by

$$p(E_t \mid p(s,c;cci)) = \prod_{i,j} p(s_j,c_i;cci)^{N_{i,j,t}} \tag{15}$$

where $p(s,c;cci)$ denotes the vector of joint probabilities of stimuli and classes; i indexes stimuli and j indexes classes, and $N_{i,j,t}$ denotes how often each combination of stimuli and classes have occurred up to trial t. In each trial, this joint probability of stimuli and classes is computed by $p(s_j,c_i;cci) = p(s_j \mid c_i;cci)p(c_i)$. We log-transformed this calculation to avoid numerical errors. For the likelihood without class-conditional independence we replace $p(s_j \mid c_i;cci)$ by $p(s_j \mid c_i;flex)$ in the above equation.

A parameter value of $\pi = 0$ reduces the model such that it makes inferences from stimulus likelihoods assuming only flexible class-conditional feature dependencies (Equation 9). A parameter value of $\pi = 1$ reduces the model such that it makes inferences assuming only class-conditional feature independence to infer the stimulus likelihoods (Equation 11). Intermediate parameter values of $0 < \pi < 1$ lead to posterior class probabilities based on a mixture of the two estimates. Note that for both $\pi = 0$ and $\pi = 1$ the posterior structural belief w equals the prior structural belief π throughout learning. For intermediate values of π, the posterior structural belief gets updated in light of the observed data. Depending on whether the environment obeys class-conditional independence, the posterior estimate of w shifts toward 1 or 0. In the two environments considered here, the posterior structural belief w decreases over time, because the structure of the environment is at variance with the assumption that features are class-conditionally independent. As experience accumulates, w gradually reduces the influence of the class predicted by class-conditional independence in favor of the class predicted by flexible conditional dependencies.

3.1.4 Differences between flexible and class-conditional independence learning

Figure 6 shows the differences between two variants of the DISC-LM, given a conservatism parameter of $\delta = 1$: one variant with a prior structural belief of $\pi = 0$ and one variant with $\pi = 1$. Imagine that the two variants of the DISC-LM experience the same five stimuli and feedback about the true classes. Both models start with flat prior beliefs. White panels in Figure 6 refer to the learning of the DISC-LM with no prior structural assumption, i.e. $\pi = 0$ (denoted FLEX for learning according to flexible dependencies); gray panels refer to the DISC-LM with assumptions of class-conditional independence, i.e. $\pi = 1$ (denoted CCI for learning according to class-conditional independence).

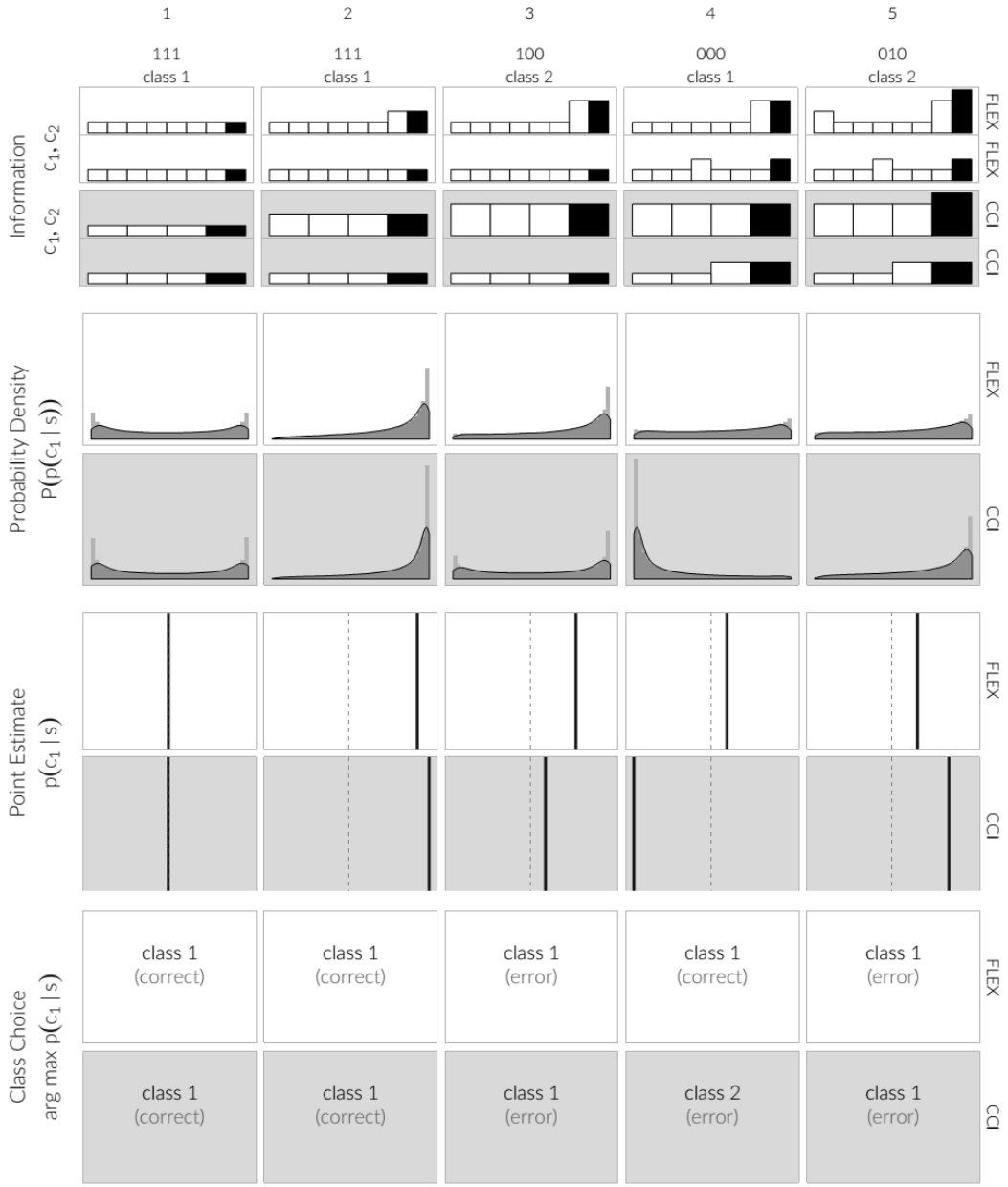

Figure 6. Model behavior. Illustration of the difference between the dependence/independence structure and category-learning model (DISC-LM) with $\pi = 0$ (denoted FLEX for flexible class-conditional feature dependencies: white panels) and $\pi = 1$ (denoted CCI for class-conditional feature independence: gray panels). The development of four properties of the models when both models learn the same sequence of stimuli and classes (left to right) is displayed. *Information:* Black bars = how often class 1 appeared. White bars, FLEX = how often each stimulus, conditional on the class, appeared. White bars, CCI = how often the marginal feature values, given the classes, appeared. *Probability density:* Posterior distribution of the probability of the class given the stimulus. Left-peaked = density favors class 1. Right-peaked = density favors class 2. *Probability point estimate:* Posterior mean of the density displayed above. Left = class 1, right = class 2, middle = guessing. *Class choice:* Decision after arg max response rule. x = wrong choice, y = correct choice. See the text for details.

The first row (information) compares how much information the FLEX and CCI variants of the DISC-LM need to extract from the environment to learn the necessary parameter. This can be interpreted as the complexity of the estimation process. The FLEX model needs information about the classes c_1 and c_2 (the black bars) to infer the class base rate. To infer the stimulus likelihoods, it needs information about seven stimuli given class 1, and seven stimuli given class 2 (the seven white bars in the white panels in the figure; the likelihood of the eighth stimulus is implied because all likelihoods for one class sum to 1). The CCI model also needs the class information, but it needs only information about the three marginal features given class 1, and another three given class 2 (the white bars in the gray panels). Thus, the FLEX version of the DISC-LM needs to extract more information from the environment compared to the CCI version.

Regarding inference from limited data, consider how certain the models are about the class given little data, illustrated in the row probability density in Figure 6. A larger variance of the posterior density (i.e., the posterior density of the probability that the current stimulus belongs to class 1) means more uncertainty about the class membership. Although both models show high uncertainty in the first trial, the class-conditional independence (CCI) model's posterior density estimate is rather peaked after a few trials, compared to the flexible independence (FLEX) model's density.

The point estimates (posterior means) of the posterior probability that the current stimulus belongs to class 1 show a similar pattern (as shown in the row probability point estimate). The estimates of the CCI version of the DISC-LM tend to be closer to the extremes, given little learning data, compared to the estimates of the FLEX version, which are closer to .50 (compare the white and gray panel for trials 4 and 5).

The bottom row of Figure 6 (class choice) shows the binarized class decisions predicted by the CCI and FLEX model. We binarized by an *arg max* rule (i.e., the model aloways picks the more likely class; see Equation 16 below). In the first trials, the models' choices are random. In trial 4, the models' decisions show how our task environment leads the CCI model astray. Based on previously observing stimulus 111 and class 1, the CCI model infers that the marginal feature value 1 points toward class 1. When the previously unseen stimulus 000 appears, it decides for the opposite class, class 2. The FLEX model, on the other hand, decides for the previously more frequent class when it observes a stimulus for the first time. It experienced class 1 more than class 2 instances and therefore chooses the former.

3.1.5 Summary

Our DISC-LM learns (a) whether the environmental structure corresponds to class-conditional independence and (b) the necessary parameters for making classification decisions. If the model parameter $\pi = 1$, the model incorporates the assumption of class-conditional independence (Equation 11). Setting this structural belief parameter to $\pi = 0$ incorporates flexible feature dependencies (Equation 9). The DISC-LM with a structural belief parameter $0 < \pi < 1$ weights the two posterior class probabilities by the fit between the observed data and the assumption about class-conditional feature dependencies.

4 Study 1 and 2: Simulations

The key question of our simulation study was how the prior structural belief π of the DISC-LM influences its behavior in our environments. Critical stimuli are those for which the class decision computed using class-conditional feature independence differs from the class decision computed using the true stimulus likelihoods. We simulated the time course of category learning in both task environments.

4.1 Stimuli

Since our focus is on early learning behavior, we generated a sequence of 50 stimuli and their respective classes by randomly drawing from the frequency distribution shown in Table 3. We generated 30 sequences of 50 stimuli each. These sequences corresponded exactly to the first 50 stimuli that participants classified in Experiment 1. We repeated this process with 50 stimuli from Environment 2 (Table 4), corresponding to the first stimuli in Experiment 2. (See below for details on human learning behavior.)

4.2 Method and Parameters

We derived the posterior probability that the current stimulus belongs to class 1 and the posterior structural belief parameter using Monte Carlo simulations. To minimize the influence of Monte Carlo error, both posterior point estimates were rounded to the fourth digit. We binarized the posterior point estimate of the probability that the stimulus belongs to class 1 by an arg max response rule:

$$\text{class choice} = \begin{cases} \text{class 1} & \text{if } \hat{p}(c_1 \mid s) > .5 \\ \text{class 2} & \text{if } \hat{p}(c_1 \mid s) < .5 \\ \text{random} & \text{otherwise} \end{cases} \tag{16}$$

where c_1 is class 1 and s the stimulus.[11]

To examine model behavior we systematically varied the prior belief in class-conditional independence

[11] In modeling, there are three primary reasons for using the deterministic arg max choice rule. First, our research focus is on comparing the model predictions with respect to the parameter π, i.e., the initial belief in class-conditional independence. A probabilistic choice rule could improve the absolute fit of the model but leave the *relative* performance depending on π unaffected. A logistic transformation of the class 1 probability, such as a softmax response rule (Wills & Kruschke, 2008), shifts the posterior probabilities toward .50 but does not shift them beyond this threshold, such as from .75 to .25. Remember that our task involves four critical stimuli for which class-conditional independence predicts one class and flexible dependencies predict the opposite class. We are interested exactly in whether the response switches from below .50 to above .50. Therefore, a probabilistic response rule would add model complexity (adding another parameter) without adding value to answer our question. Second, probabilistic choice rules require aggregating data over individuals or trials (which is common practice, e.g., D. Friedman et al., 1995). Aggregating over trials assumes little or no co-variance of choices over time (Hannan, 1985). However, learning data is characterized by the very dependencies of decisions over time. Therefore, time aggregation would not do justice to our data. Aggregating over individuals is also not possible because people varied greatly in their learning speed (i.e., the number of trials they needed to hit the learning criterion in our task; see the experimental results). The third reason for the arg max rule is pragmatic. The deterministic choice makes it easiest to illustrate how the parameter π changes the DISC-LM's performance.

π of the DISC-LM. We ran simulations for values of $\pi \in \{0, 0.90, 0.99, 0.999999, 1\}$. A structural belief value of $\pi = 0$ corresponds to a classifier with no prior structural beliefs and a value of $\pi = 1$ to a classifier that always assumes that features are class-conditionally independent (i.e., the naïve Bayes model). The conservatism parameter was set to $\delta = 1$ in all simulations (A.2.6 shows how different values of δ affect model predictions).

4.3 Results

The key question was how the DISC-LM learns the critical and uncritical stimuli depending on its belief in class-conditional independence. Figure 7 shows the results in terms of model learning curves of the true class of each stimulus, and the learning of the true feature dependency structure, in Environment 1. Figure 8 shows the corresponding results for Environment 2.

Learning curves. With stronger prior beliefs in class-conditional independence π learning slows down, but only for the critical stimuli 000, 001, 010, and 100. Stimulus 111 remains largely unaffected by the value of π, as the top panel of Figure 7 shows. Note that the curves are not completely smooth, which is due to the randomness of the sequences drawn. For the probabilistic environment, this holds as well.

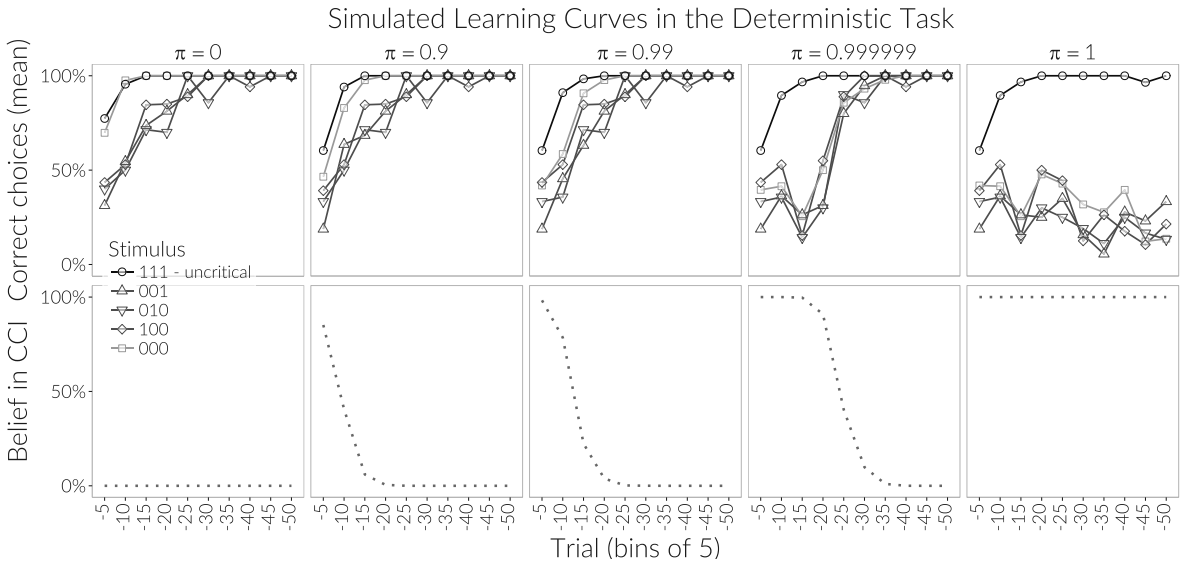

Figure 7. Model learning (environment 1). How quickly the DISC-LM learns to classify the stimuli in Environment 1 depends on the prior beliefs in class-conditional independence (CCI), π. Stronger prior beliefs in class-conditional independence result in slower learning, but only for the critical stimuli. Also note that the leftmost model ($\pi = 0$) performs above chance for stimuli 000 and 111 in the first bin. This is because the DISC-LM with $\pi = 0$ infers the class of stimuli 000 and 111 correctly from the class base rate within five trials. *Belief in CCI:* The belief in CCI decreases with experience in the environment, for prior belief values of $0 < \pi < 1$ (higher values represent stronger beliefs). Note: The x-axis shows the trials in bins of five while keeping the presentation order of stimuli.

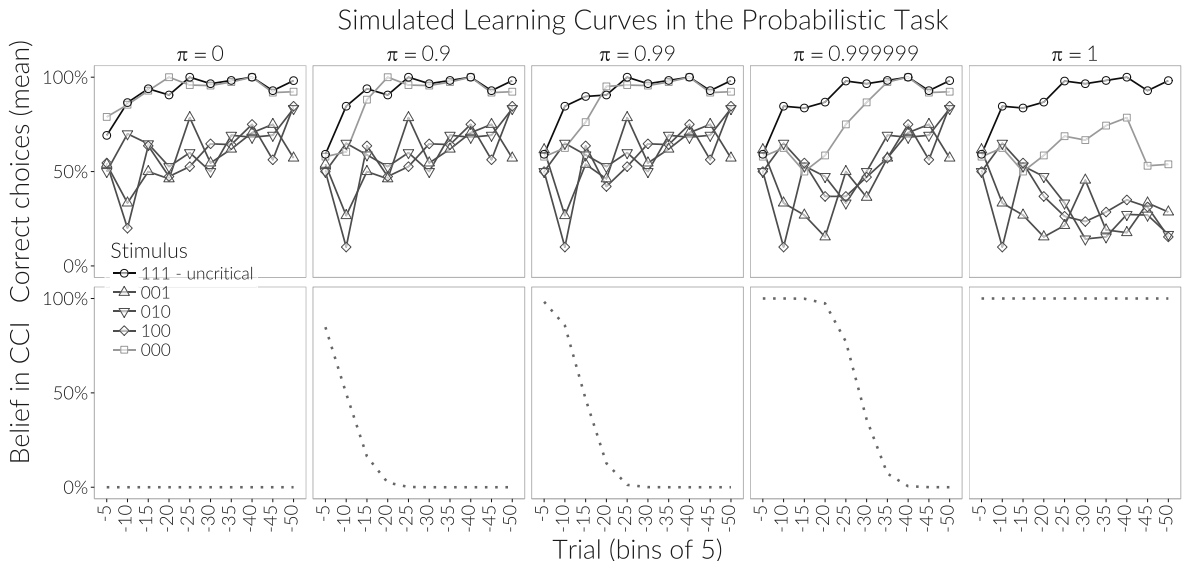

Figure 8. Model learning (environment 2). How quickly the DISC-LM learns to select the most likely class for all the stimuli in the probabilistic Environment 2 depends on its prior beliefs in class-conditional independence (CCI), π, and on the stimulus. The qualitative pattern is similar to model behavior in the deterministic Environment 1. Note, however, that the maximum performance is lower due to the probabilistic nature of the task. *Belief in CCI:* The belief in CCI decreases with experience in the environment, for prior belief values of $0 < \pi < 1$ (higher values represent stronger beliefs). Note: The x-axis shows the trials in bins of five while keeping the presentation order of stimuli.

The DISC-LM without structural assumptions ($\pi = 0$) quickly learns to correctly classify all stimuli, in both environments. Due to the probabilistic nature of Environment 2 (Figure 8), the maximum performance the model can achieve is slightly lower (see Table 4). The differences in learning speed for the different feature configurations, in both environments, reflect the unequal frequencies of the stimuli (see Table 3, and Table 4). Stimuli 111 and 000 are more frequent and therefore learned fastest, as opposed to 001, 010, and 100 which are equally but less frequent.

The DISC-LM with $\pi = 1$ believes in class-conditional feature independence throughout learning. This model learns quickly to correctly classify the uncritical stimulus 111, in both environments, but fails to learn the critical stimuli. Even with an infinite amount of experience this model will—in this particular environment—never learn, because the model's fixed belief in class-conditional independence prevents it from learning the true dependencies among features.

The DISC-LM with high (but nondeterministic) prior structural beliefs ($\pi = .9, .99, .999999$) learns the uncritical stimulus as quickly as the DISC-LM that assumes flexible dependencies ($\pi = 0$). However, high values of π result in slower learning of the critical stimuli. The stronger the prior belief in class-conditional independence, the greater the slowing in learning. Values lower than $\pi = .9$ yield results similar to $\pi = 0$ because the belief in class-conditional independence is quickly overridden by the contradictory learning data.

Learning environmental structure. In addition to the environmental probabilities, the DISC-LM

learns the overall statistical structure of the environment, in particular whether features are class-conditionally independent. A stronger prior belief in class-conditional independence π slows down learning the dependency structure, which is shown in the bottom row of Figures 7 and 8. As soon as the model has updated the structural assumption about the environment (i.e., realized that class-conditional independence does not hold), it starts to improve on the classification of the critical stimuli in the top row of the figure.

4.4 Summary

With strong prior structural belief in class-conditional independence, learning of the DISC-LM is influenced asymmetrically, in both environments. With stronger prior structural beliefs, the critical stimuli 000, 100, 010, and 001 are learned more slowly, but learning of stimulus 111 is not impaired (Figures 7 and 8). This is true in the deterministic and probabilistic environment. These results lend support to the predictions summarized in Table 5.

Table 5

Predictions for the stimuli in our task from the simulation study.

Simulation Result	Description
Superiority of 111	Stimulus 111 is learned most quickly, largely independent of the prior belief in class-conditional independence π
Initial slowing of 000	Learning of stimulus 000 is slower in the first trials the stronger the prior belief in class-conditional independence
Slowing of 001, 010, 100	Learning to classify stimuli is slowed down uniformly with stronger prior beliefs in class-conditional independence
Similarity of 000 and 111	The model with $\pi = 0$ predicts that stimuli 000 and 111 are learned almost equally fast

5 Study 3: Experiment 1 — Deterministic Task

The goal of Experiment 1 was to investigate whether human learners treat features as class-conditionally independent early in learning. Our experiments used a supervised trial-by-trial learning paradigm (e.g., Ashby & Maddox, 1992) adapted from previous studies (e.g., Nelson et al., 2010; Meder & Nelson, 2012). Experiment 1 was based on the deterministic task environment shown in Table 3.

5.1 Participants

Thirty people (mean age 23.8 years, range 19 to 33 years, 67% female) participated; remuneration was 12 euros. We recruited via the Center for Adaptive Behavior and Cognition at the Max Planck Institute for Human Development in Berlin, Germany. Data were collected from September to December 2012 at the Center; the experiment was conducted in accordance with the ethical and data protection guidelines there.

5.2 Materials and procedure

Participants classified "plankton" stimuli differing in eye, claw, and tail appearance (three binary features) into species A and species B (binary class). Figure 9 illustrates the material. Each plankton specimen corresponded to one feature configuration in Table 3. The assignment of physical features and class labels was randomized across participants.

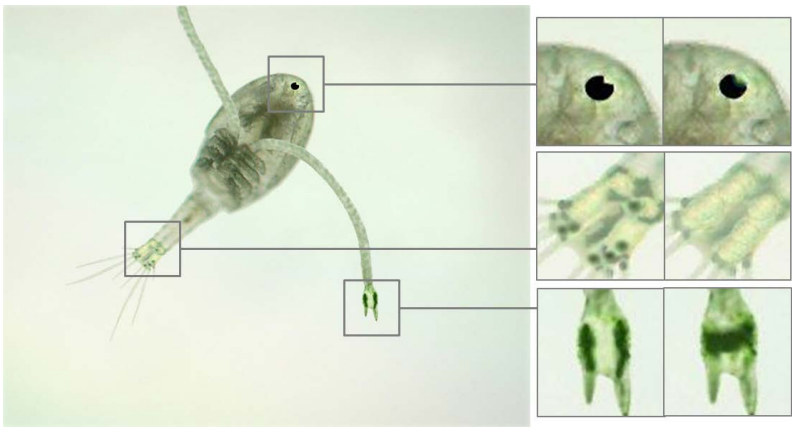

Figure 9. Sample stimulus used in Experiments 1 and 2 (from Nelson, 2010). In each trial, participants saw and classified one plankton specimen (left) on the basis of three binary features. The gray boxes and magnification of the three features (right) are for illustrative purposes only.

First, participants were familiarized with the plankton stimuli, their features, and the feature locations. Then they began learning. In each trial, participants saw a plankton specimen drawn randomly from the task distribution (Table 3). They classified it by pressing the left-arrow or right-arrow key (for species A or B). Subsequently, the true species appeared at the center of the screen (as the letter "A" or "B"), overlaid with a smile emoticon after a correct decision or a frown emoticon otherwise. Learning was self-paced; participants proceeded by pressing the up-arrow key. They were instructed to always choose the most likely class, which in a deterministic task means to correctly classify the stimulus. The presentation of stimuli ended when participants reached a criterion defined as (a) having made at most four classification errors over the last 200 trials (98% of 200 correct), and (b) having chosen the most likely category for the last five times that each individual stimulus appeared within the random sequence of stimuli.

After 15 learning trials participants saw "frequently asked questions," which, among other things, reminded them to always pick the most likely class and informed them that it usually takes 300—400 trials to reach criterion performance. At regular intervals (every 100 trials, from trial 200 onward), participants were informed about their current accuracy and the maximum possible accuracy in the task and reminded to learn all features (see A.3.1 for the wording).

5.3 Behavioral results

All participants reached the learning criterion in 200 to 798 trials (median = 338, mean = 381, SD = 155).

Classification errors. We expected that if participants assumed class-conditional independence in our first environment, we would see more classification errors for the critical stimuli (000, 001, 010, 100) than for the uncritical stimuli, overall. We derived error rates separately for each stimulus, because stimuli were not equally frequent (see Table 3). Figure 10a shows participants' classification errors. The critical stimuli, for which assuming class-conditional independence entails a different class decision than the true environmental statistics, were more frequently misclassified than the uncritical stimulus, over the whole course of learning.[12] This finding is consistent with the idea that people treated features as class-conditionally independent and classified stimuli accordingly.

Aggregating errors over time and individuals ignores both interpersonal variability and temporal dynamics (e.g., W. K. Estes & Maddox, 2005). The development of errors over time is key to our hypotheses. The fact that all participants achieved criterion performance indicates that they did not treat features as class-conditionally independent over the whole course of learning; otherwise they would have continued to misclassify the critical stimuli and failed to meet the learning criterion. Our hypothesis, formalized in the DISC-LM, is that learners initially assume class-conditional independence, which should slow down learning in the early trials, but not at later stages, when a sufficient amount of data contradicting this assumption has been observed. The next set of results show the temporal learning dynamics at the aggregate level, and thereafter the modeling of the individual learning dynamics.

Learning curves. The simulations showed that higher prior beliefs in class-conditional independence impair learning asymmetrically for the critical stimuli, but not for stimulus 111 (Figure 7). The DISC-LM predicted a superiority of stimulus 111 for all parameter values of the prior belief π in class-conditional independence. The observed stimulus-wise learning curves in Figure 10b reveal that people learned feature configuration 111 the quickest. For stimulus 000, on the other hand, only values of $\pi > 0$ predicted a learning impairment relative to stimulus 111. The observed data show slower learning of 000 than 111. This contradicts the simulated learning curve of the DISC-LM model with $\pi = 0$—the model without structural priors—according to which stimulus 000 should be learned as fast as 111. Moreover, the models with high prior beliefs (above .90) in class-conditional independence predicted that the remaining stimuli (001, 010, 100) would be learned after an initial phase of stagnation. Participants' learning curves also show this pattern. This provides evidence for an initial structural belief consistent with a high prior on class-conditional independence of features.

Because aggregated learning curves ignore inter-individual variation, we also modeled the individual learning trajectories. Some participants may learn according to a more flexible conditional feature independence (e.g., the participant who reached the learning criterion in the minimum possible of 200

[12] This analysis used all trials, i.e., including the last 200 trials for which our learning criterion enforced 98% correct choices, because excluding the last 200 trials resulted in 19 (of 30) participants being left with fewer than 20 learning trials for one or more of the five stimuli. If we compute the median of the proportion of errors after excluding the last 200 trials, the qualitative result is unchanged, i.e., fewest errors for the uncritical stimulus ($111 < 000 \approx 100 \approx 010 \approx 001$ with median error rates .09, .22, .21, .22, .20 respectively).

trials), while others may hold stronger beliefs in conditional feature independence.

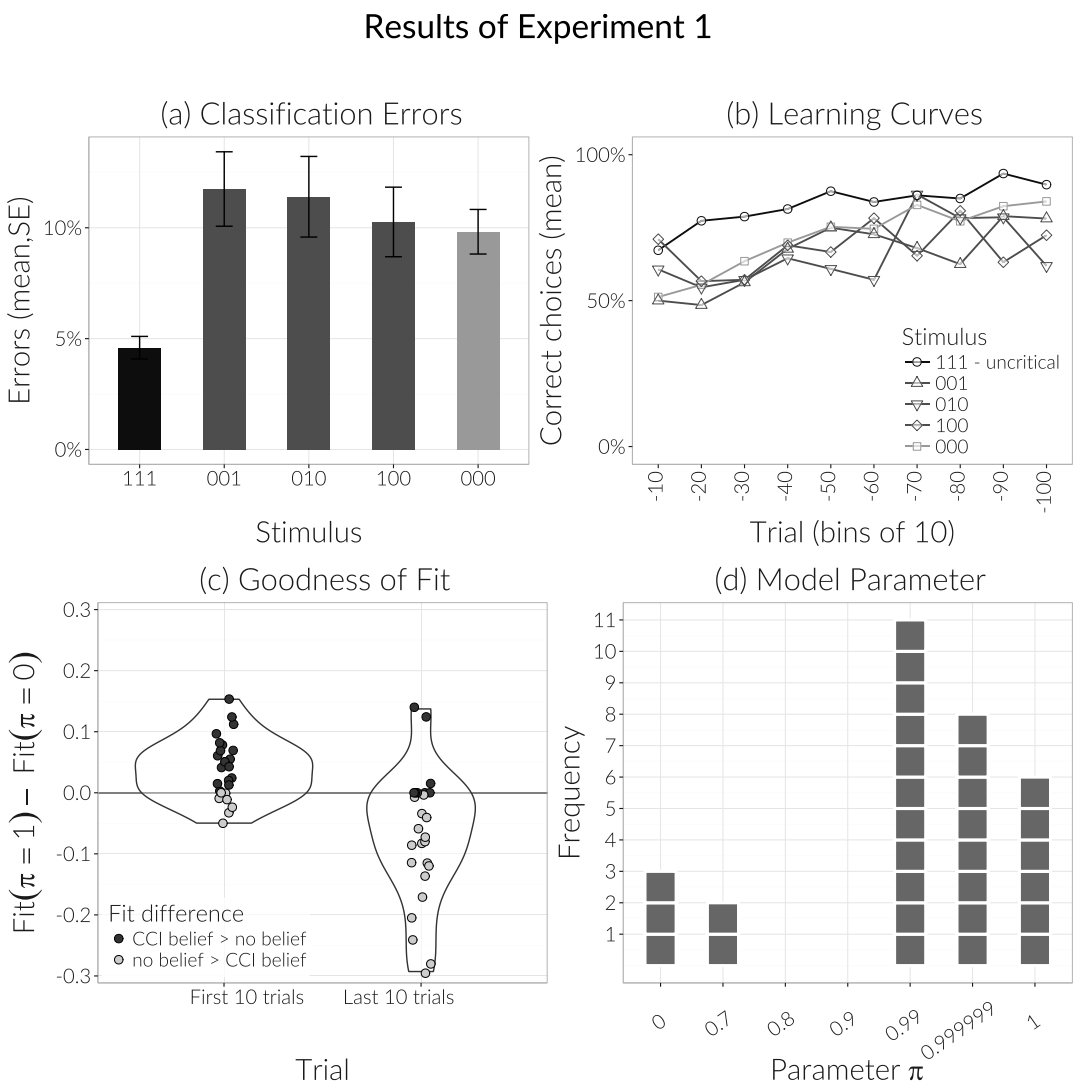

Figure 10. Results of Experiment 1. **(a)** Average error rates by stimulus. Error bars represent standard errors, bootstrapped with 100 replications. Participants' error rates were lowest for the uncritical stimulus (111) and higher for the four critical stimuli (000, 001, 010, 100). We used all trials to compute the errors. **(b)** Proportion of correct choices during the first 100 trials. Dots represent correct (i.e., most likely) class choices per stimulus, averaged within bins of width 10. The curves show that the uncritical stimulus (111) was learned faster than the critical stimuli, for which the class prediction using class-conditional independence contradicts the prediction using flexible dependencies (in the limit). Among the latter, learning stimulus 000 was easiest, but still harder than 111. **(c)** In the first 10 trials, a model assuming class-conditional independence ($\pi = 1$) shows a higher fit than a model without the assumption ($\pi = 0$) for most participants; however, this reverses for the last 10 trials. **(d)** Distribution of the prior structural belief parameter π when predicting individual choices of participants ($N = 30$). A value of $\pi = 0$ means flexible conditional feature dependencies; a value of 1 means a strong fixed prior belief in class-conditional independence. Models with a high prior on class-conditional independence best account for the majority of participants in Experiment 1. Note: Models were compared by mean squared error (MSE) of their one-trial-ahead predictions on trials $t = 2$ to $t = \max(t) - 200$. We defined the thereby excluded participant as belonging to $\pi = 0$ (see main text).

5.4 Modeling results

Data preprocessing. Because the DISC-LM assumes uniform prior distributions, it predicts equal class probabilities in the first trial, $p(c_1) \approx p(c_2) \approx .50$, except for small Monte Carlo errors. We tested if participants' observed first choices deviated from randomness and found no difference (19 of 30 class a choices in the first trial, exact binomial test for equal proportions: $p = .21$). To ensure that the Monte Carlo error did not distort the comparison of model and human behavior, we predicted participants' decisions only starting from trial 2. Further, we used the observed decisions only up to $T - 200$ where T is the participant's last trial, because in the very last 200 trials our learning criterion enforced 98% correct choices. This excluded one participant who needed exactly 200 trials, for whom we assumed no prior belief in class-conditional independence, i.e. $\pi = 0$. This left 29 participants for our analyses.

Prediction generation. To investigate at a more fine-grained level whether class-conditional feature independence is a default assumption in human category learning, we were particularly interested in the parameter π of the DISC-LM (i.e., the prior belief in class-conditional feature independence). We applied individual parameter selection through one-trial-ahead prediction, using mean squared error (MSE) as criterion for the quality of the prediction.[13] The MSE was our measure of choice because it emerged as the best measure in a parameter recovery simulation, compared to mean absolute error and likelihood-based measures.

We modeled the decision of each learner in each trial with the DISC-LM. The classification probabilities were derived from Monte Carlo simulations using a grid of a priori fixed parameter values. This grid included $\pi \in \{0, 0.7, 0.8, 0.9, 0.99, 0.999999, 1\}$; for δ (conservatism) we used $\delta \in \{1, 2, 3, 4, 7, 12, 20, 33, 55, 90, 148\}$. We used a finer grid resolution for π close to 1 because the model predictions in the lower grid regions were rather similar to each other, ceteris paribus. Each prediction by models with values of $\pi \leq 0.6$ differed by less than 1 percentage point from predictions by a model with $\pi = 0$, when comparing trials 2 to 100. By contrast, changing π from .9 to .99 resulted in a substantial difference in the predicted point estimates of the class membership. We rounded the predictions to the fourth digit.

Goodness of fit. If class-conditional independence is used as the default assumption by our participants, a model incorporating feature independence should outperform a model with no such assumptions on the early but not late data. We compared the two most extreme settings of the DISC-LM on how closely they resemble people's first and last 10 classification decisions. We fixed the model belief in class-conditional independence π to $\pi = 0$ or $\pi = 1$. We computed the difference in fit in terms of $1 - MSE$ between these parameter settings for each subject for the early and the late trials (given individually adjusted δ values). The fit differences ($f_{\pi=1} - f_{\pi=0}$) in Figure 10c show that the class-conditional-independence DISC-LM accounts better for the data early in learning (positive fit difference; mean $= 0.04$, $t(25) = 4.02$, $p = .0005$) but the class-conditional-independence DISL-LM performs worse for the late learning data (negative fit difference; mean $= -0.07$, $t(25) = -3.40$, $p = 0.003$).[14]

[13] For each participant, the individual MSE was computed as $MSE = \frac{1}{T}\sum_t (c_t - \hat{p}_t)^2$, where t indexes trials, T is the number of trials used for parameter selection, c_t denotes the participant's choice for trial t, and $\hat{p}_t$ denotes the predicted probability for the class. This was the simulated expected value of the classification beliefs for each trial (see A.2.4 for details).

[14] The reduced number of degrees of freedom ($df = 25$ instead of 28 with 29 participants) result from the fact that some

Model accuracy. To analyze the individual choices over the whole learning duration, we used the full parameter grid to obtain the parameter combination for π and δ that jointly minimized the MSE between observed choices and model predictions. We counted the excluded participant (who had too little learning data) as a learner who did *not* use class-conditional independence, with values of $\pi = 0$ and $\delta = 1$. Given the resulting parameter values, the model's accuracy was 81.77%, averaged across the 29 participants.[15] By contrast, a DISC-LM without independence assumptions that enforces flexible dependencies ($\pi = 0$ and best-fitting values of δ) is less accurate, reaching only 77.42% accuracy; with a mean difference between those models of .04 ($t(28) = 3.58$, $p = .0012$).

Initial beliefs in class-conditional independence. We hypothesized that humans start classification learning with an initial belief that features are class-conditionally independent, that is with high values of π. The data strongly bear out this expectation. Of our 30 participants, 25 were best fit by a model with prior belief in class-conditional independence of at least 0.99 (Figure 10; Appendix A.2.8 gives results for the conservatism parameter). This suggests that class-conditional independence is used by the majority of participants early in classification learning. Only a minority (3 of 30) did not have prior beliefs in class-conditional independence (i.e., $\pi = 0$ was the best fitting parameter).

Summary. Both the analyses of the classification errors for the different stimuli on the aggregate level, and the analyses on the individual level are consistent with the idea that class-conditional independence serves as a default assumption in classification learning. The classification errors and different learning curves are in line with a model that assumes a strong initial belief in conditional independence. When fitting the π parameter of the DISC-LM to the learning data, for most participants a high value of π accounted best for the data. We next investigated the assumption of class-conditional independence in the context of a probabilistic classification task.

6 Study 4: Experiment 2 — Probabilistic Task

6.1 Participants

A total of 39 people participated. Ten had to be excluded (8 who did not reach the learning criterion in 120 min, 2 due to a computer crash), leaving us with 29 participants (mean age 24.8 years, range 18 to 35 years; 79% female). They were paid 12 euros. Data were gathered from April to June 2013 at the same laboratory as in Experiment 1.

6.2 Materials and procedure

The materials and procedure were almost identical to those in Experiment 1, with the difference that the stimuli were drawn from the probabilistic task environment (Table 4). The correct (most likely) choices given the stimuli corresponded to Experiment 1, but the maximum achievable accuracy was 88% (instead of 100% as in Experiment 1).

people learned too quickly, in less than 20 (i.e., 10 early, 10 late) trials.

[15] Accuracy was defined as the number of trials in which observations corresponded to the model predictions after binarizing the probabilistic predictions by an arg max response rule (Equation 16).

6.3 Behavioral results

Participants reached criterion performance in 212 to 1,156 trials (median = 627, mean = 620, $SD = 280$), which is slower compared to Experiment 1, $t(43) = 4.03$, $p = .0003$, Cohen's $d = 1.06$.[16] The slower learning in the probabilistic compared to the deterministic environment corresponds to previous findings (e.g., Little and Lewandowsky, 2009; Nosofsky and Stanton, 2005; Mehta and Williams, 2002; but see Juslin et al., 2003; Seger and Cincotta, 2005).

Classification errors. As in Experiment 1, we computed the proportion of errors separately for each stimulus, aggregating over time and participants.[17] Again participants made more errors when classifying the critical stimuli, compared to the uncritical stimulus (Figure 11a)[18] (The critical stimuli were those for which assuming class-conditional independence results in diverging class choices than when assuming flexible dependencies.) The relatively small number of errors for the critical stimulus 000 can be explained considering Table 4, which shows that, in this environment, a classifier with class-conditional independence predicts the correct class of 000 with a rather high probability ($p(c_1 \mid 000; cci) = .48$).

Learning curves. Figure 11b shows the stimulus-wise learning curves, aggregated over participants. The pattern corroborates the results of Experiment 1: The easiest item was the uncritical stimulus 111; the critical stimulus 000 was more difficult, at least in the beginning. This pattern is predicted only by models with beliefs in class-conditional independence, and not by a model that a priori assumes flexible conditional feature dependencies ($\pi = 0$). Critical stimuli 001, 010, and 001 were the most difficult ones, consistent with strong beliefs in class-conditional independence. Again, the data are at variance with the pattern predicted by the DISC-LM with $\pi = 0$, according to which stimuli 000 and 111 should be learned equally quickly. These findings are in line with the results of Experiment 1, supporting the hypothesis that human learners initially treat features as class-conditionally independent.

[16] With Welch–Satterthwaite correction for variance inhomogeneity.

[17] In a probabilistic task, the term "error" needs to be distinguished from (in)accuracy, because for a given stimulus the true class need not equal the most likely class. Classification errors result from choosing a class other than the *most likely* class, while inaccuracies result from choosing a class other than the *true* class. For instance, the most likely class for stimulus 111 is class a, because $p(a|111) = .89$. A class b choice constitutes a classification error, whereas class a choices are correct, irrespective of whether in that particular trial the stimulus belonged to category a. Consequently, error rates range from 0% to 100%, whereas the expected average accuracy ranges from 0 to 88 of 100 trials in our task.

[18] Again, our analysis used all trials, i.e., including the last 200 trials for which our learning criterion enforced 98% correct choices, because excluding the last 200 trials resulted in 9 (of 29) participants with fewer than 20 choices for at least one stimulus type. When excluding the last 200 trials, the order of the median error rates corresponds to using all trials: The order is $111 < 000 < 100 < 010 \approx 001$ (.09, .15, .33, .37, .39 respectively).

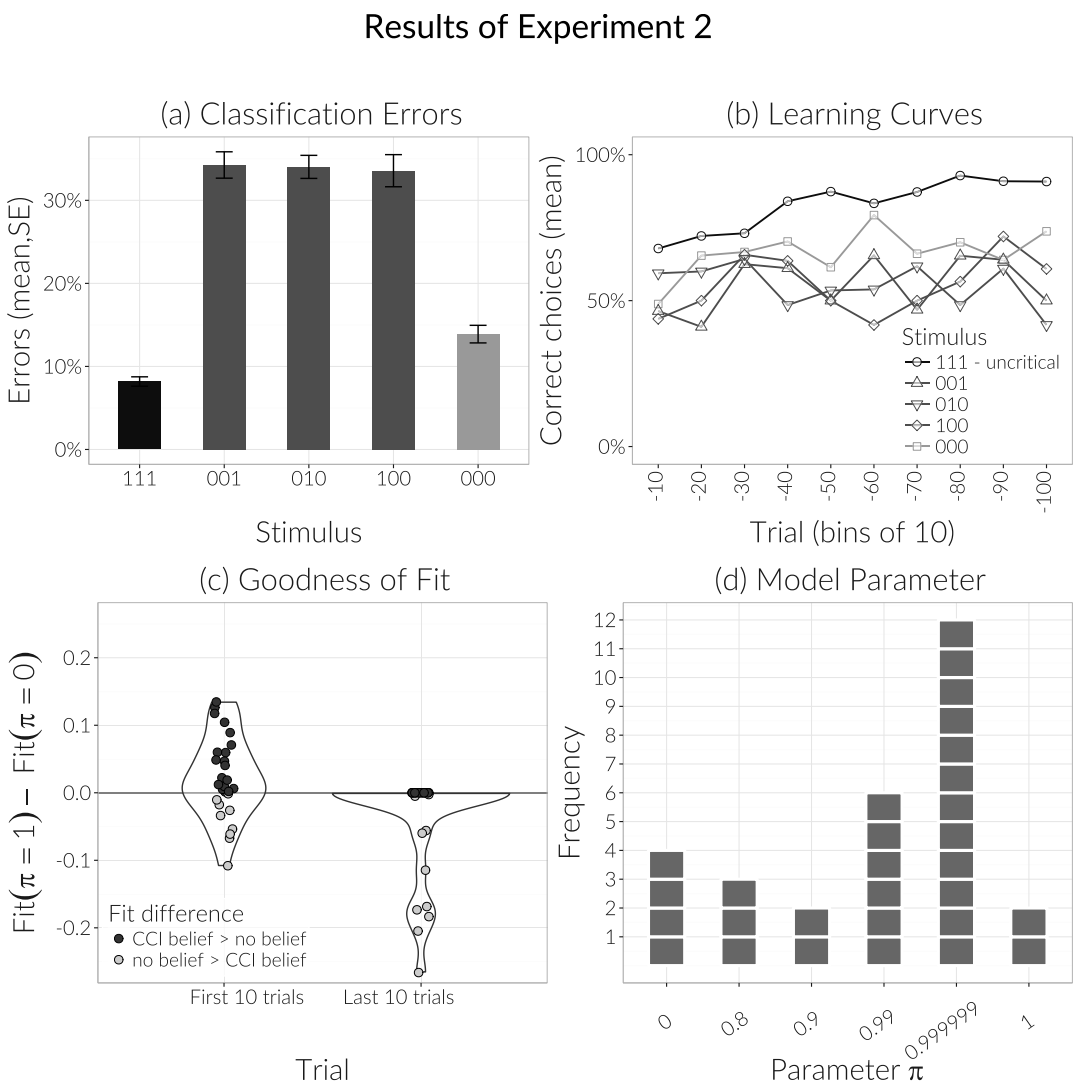

Figure 11. Results of Experiment 2. **(a)** Classification mistakes by stimulus. **(b)** Classification improvement in the first 100 trials. Dots represent most-likely-class choices per stimulus, averaged within bins of width 10. The curves show that participants improved fastest for the uncritical stimulus (111). For critical stimuli, improvement was slower. Among them, learning stimulus 000 was easiest, but still harder than 111. **(c)** In the first 10 trials, a model assuming class-conditional independence ($\pi = 1$) shows a higher fit than a model without the assumption ($\pi = 0$) for most participants; however, this reverses for the last 10 trials. **(d)** Distribution of best predicting values for the parameter π ($N = 29$). This parameter reflects the model's prior belief in a class-conditionally independent task structure. Values of $\pi = 0$ denote no belief; values of 1 denote the strongest structural belief. The model with high belief values predicts most participants best in Experiment 2. Note: Model accuracy measure was mean squared error (MSE); we used trials $t = 2$ to $t = \max(t) - 200$ to obtain the values.

6.4 Modeling results

We used all trials except the first trial and the final 200 trials to examine which values of the prior belief in class-conditional feature independence π best predicted participants' decisions.[19] We derived predictions using Monte Carlo simulations of the DISC-LM for different values of π and δ. The parameter grids consisted of $\pi \in \{0, 0.8, 0.9, 0.99, 0.999999, 1\}$, and $\delta \in \{1, 2, 3, 4, 7, 12, 20, 33, 55, 90, 148\}$.[20]

Goodness of fit. Comparing how well the extreme versions of the model (the DISC-LM that can learn flexible dependencies and the one that always uses class-conditional independence) predict early and late learning behavior shows that the class-conditional-independence DISC-LM performs slightly better for early choices but worse for later data. The differences in fit ($f_{\pi=1} - f_{\pi=0}$) are positive for the early and negative for the late trials.[21] Figure 11c show a trend favoring the class-conditional-independence model in the first ten trials (slightly positive fit difference; mean $= 0.02$, $t(27) = 1.86$, $p = 0.07$); while in the last ten trials the class-conditional-independence DISC-LM performs worse (negative fit difference; mean $= -0.04$, $t(27) = -2.93$, $p = 0.006$).[22]

Model accuracy. We obtained the best parameter values by individual trial-by-trial predictive fitting as described for Experiment 1. There was one tie where two π values resulted in equal MSE scores. We conservatively selected the lower π value for a lower belief in class-conditional independence. The model's accuracy for the resulting parameter values, averaged across the 29 participants, was 84.12%. A model that does not believe in class-conditional independence ($\pi = 0$ and best-fitting δ values) has 82.83% accuracy, which yields a mean difference of .01 ($t(29) = 1.44$, $p = .16$). Here, only the qualitative direction indicates that a $\pi = 0$ model describes behavior less well. There are two reasons for the rather small accuracy difference in Experiment 2. First, in about 38 of 100 trials, when the uncritical stimulus is drawn, the models make the same choice prediction irrespective of π (see Table 4). Re-computing the accuracy without stimulus 111, yields a bigger mean accuracy difference of 79.34% and 77.58% (mean difference $= .02$, $t(28) = 1.80$, $p = .09$) between the best-fitting-parameter and the flexible-dependency DISC-LM. The second reason is that after the structural belief parameter w is updated to $w = 0$ (which happens quickly after ca. 50 trials, see the lower panel in Figure 8) both models become indistinguishable in the later learning trials. Experiment 2 involves many more of the late learning trials (the median number of trials lies beyond 600), making it harder to discriminate the models based on the average fit across all learning trials.

Initial beliefs in class-conditional independence. If human learners are initially guided by high prior beliefs in class-conditional independence, this should be reflected in high parameter values of π. In line with this, we obtained values of $\pi \geq 0.9$ for the majority of participants (right panel (d) in Figure 11; results for the parameter δ are shown in A.2.8). As in Experiment 1, few participants (4 of 29) were best

[19] We tested whether we excluded informative data by not using the first trial and found no evidence for this: 12 of 29 participants selected class *a* in the first trial, exact binomial test for equal proportions: $p = .46$.

[20] The grid was determined (as in Experiment 1) by selecting parameter values such that each prediction differed by more than 1 percentage point from predictions with $\pi = 0$, across trials 2 to 100.

[21] Again, we computed the fit for $\pi = 0$ and $\pi = 1$ given individually adjusted δ values.

[22] Again, the number of subjects used for this analysis is lower than the total number (29) because one subject had less than 20 (10 early, 10 late) learning trials.

described by a fully flexible model with no beliefs in class-conditional independence. The behavior of most participants (20 of 29) was best accounted for by strong beliefs in class-conditional independence, with π values of 0.99 or higher. These results support the idea that class-conditional independence serves as a default assumption in human classification learning.

7 General Discussion

A variety of theoretical arguments and machine learning results suggest that the assumption of class-conditional independence could be very helpful in probabilistic category learning. Research to date with human subjects has not specifically focused on this question. We used computer simulations to identify statistical environments in which learners who presume class-conditional independence will make strongly different classification decisions than fully flexible learners.

With a new Bayesian learning model, the DISC-LM, we studied the kinds of inferences that would be made over the course of learning, according to whether or how strongly one initially presumes class-conditional independence. Different versions of the DISC-LM, with different prior beliefs about environmental structure, showed very different patterns of learning trajectories, especially early in learning. Importantly, the DISC-LM learns, over time, whether or not class-conditional dependence holds and adjust its beliefs and classification decisions accordingly. Based on the model behavior we derived a number of specific predictions for human learners' behavior, on the same tasks.

Experiment 1 consisted of a deterministic classification learning task designed to optimally discriminate between people who treat features as class-conditionally independent and those who do not. The task involved four critical stimuli, for which classifications derived assuming class-conditional feature independence disagreed with the class choices derived from the true feature dependencies, and one uncritical stimulus, for which the independence assumptions did not entail diverging classifications. Participants made more classification errors for the critical stimuli than for the uncritical stimulus. Participants also learned the critical stimuli more slowly compared to the uncritical stimulus. We also modeled individual choices, using the DISC-LM. Most participants' classification decisions were best predicted by versions of the model with very high prior belief in class-conditional feature independence.

Experiment 2 followed a similar rationale but used a probabilistic task, to reflect that most real-world categorization environments are not deterministic (either inherently or due to incomplete knowledge). Results replicated the first experiment: Most participants' initial classification decisions were best accounted for by a DISC-LM with a high prior belief in class-conditional independence.

Results from all analyses, across both experiments, found that models that place extremely strong (but not 100%) initial belief in class-conditional independence best account for human behavior. Note that the version of the DISC-LM with complete belief in class-conditional independence (i.e., $\pi = 1$), which does not learn to correct its initially structural assumptions if the experience contradicts them, did not capture participants' behavior; neither did a version of the model that allowed for completely flexible conditional feature interactions (i.e., $\pi = 0$) throughout learning. Although class-conditional independence performs well across many environmental structures; people can learn from experience

to overcome their structural prior beliefs when the learning input contradicts their assumptions.

We used environments with three binary features, and two binary categories. The literature includes many tasks with two or three features (e.g., Sanborn et al., 2010; Meder and Nelson, 2012; Vigo, 2013; Rehder and Burnett, 2005; but see Nosofsky, Palmeri, and McKinley, 1994); thus we had a priori reason to believe that such tasks would be learnable. In environments with more than three features, the curse of dimensionality is stronger; thus, making a simplifying initial assumption, such as class-conditional independence, would be even more important in more complex environments.

7.1 Relation to other models

As outlined in the Introduction, models based on prototypes (Reed, 1972) have been suggested to describe early learning behavior (J. D. Smith & Minda, 1998) better than exemplar-based models. The formation of prototypes in these models relates to class-conditional independence, as prototypes consist of the mean or median value of each feature dimension per class. Furthermore, such models predict that linearly separable category structures will be easier to learn, which directly relates to the class-conditional independence assumption, as shown earlier. Independence assumptions also occur in decision boundary models, which consider a geometric representation of the feature and class combinations. These models (e.g., Ashby & Maddox, 1992, 1990) state that people divide the multidimensional feature space into different regions for each class and assign a stimulus according to the region it falls into. Let us consider two classes of decision bound models, and their relation to class-conditional independence. The first is independent cue models, which assume that people form a boundary for each feature (that is, boundaries parallel to the axes of the feature space), and the intersection of these unidimensional boundaries determines category membership (e.g. Ashby & Maddox, 1990; Ashby & Gott, 1988). This approach, although it will also fail in our environments, is not equivalent to presuming class-conditional independence, because class-conditional independence integrates features. A class-conditional independence classifier, however, can learn independent cue decision bounds. The second class of decision bound models consists of linearly-separable models. Linearly separable decision bounds are obtained by combining features by an interaction-free, weighted sum. As pointed out earlier, this relates to the naïve Bayes model, which is linear in log space. For binary features the naïve Bayes model is restricted to linearly separable features (M. Jaeger, 2003; Zhang & Ling, 2001), but if the features can take more than two values this need not hold (Zhang & Ling, 2001).

The late learning behavior, where our participants learned the category structure and utilized the feature combinations, can potentially be described by various models. It may be described with an exemplar-based strategy, which can learn exclusive-OR structures such as our task environments (Nosofsky, 1992). Late learning may alternatively be described with a rule-plus-exception strategy, for example, "classify as class 1 if all features = 1, otherwise classify as class 2, except if all features = 0." As noted before, our purpose was not to develop a new categorization model, but rather to test whether participants bring a specific assumption that is justified from a computational perspective (class-conditional independence) when learning novel categorization tasks.

It should be noted that the probabilistic DISC-LM is situated at Marr's computational level (Marr, 1982) and does not make claims about the underlying cognitive information-processing steps. We designed the model to test a specific hypothesis about people's behavior and not so much the cognitive processing implementing the behavior. The insights from our studies of the DISC-LM and human learners are potentially relevant to researchers building various kinds of learning models.

The next steps in developing the DISC-LM are (a) to test the predictions that it provides about the development of the learners' beliefs about the structure of the task (lower panels of Figures 7 and 8), and (b) to translate the assumption of class-conditional independence into specific process model predictions, which entails tweaking the model such that it predicts a second, independent data dimension such as reaction times or electroencephalogram data or gaze pattern; Jarecki, Tan, and Jenny, 2016, in addition to choice predictions.

7.2 Implication for the study of strategy selection

Our findings complement studies on the strategy selection problem (how do people adapt their decision and inference strategies to the nature of the task?), which has only recently been explicitly addressed (Rieskamp & Otto, 2006; Gluth, Rieskamp, & Büchel, 2014; Glöckner & Betsch, 2008; Mata, von Helversen, & Rieskamp, 2011; Marewski & Schooler, 2011; Lieder & Griffiths, 2015). For instance, Gluth, Rieskamp, and Büchel, 2014 found evidence that in multiple-cue inference tasks people's behavior and neuronal data is best described by a model formalizing dynamic switches between decision strategies over time. Our approach shows that beliefs about the nature of the task (feature dependencies) could at least implicitly be a guiding principle by which people learn to adapt inference strategies.

7.3 Implications for knowledge-specific learning

Our data also informs the literature on the interaction between the context of a task and the specific structural knowledge people apply. For example, participants in an experiment by Wattenmaker et al., 1986 expected to learn a linearly separable categorization structure when the cover story of a person-classification task was such that the features of one category coincided with aspects of one personality trait. However, participants did not expect linear separability when the features associated with one category belonged to different character traits. Thus, the underlying knowledge structure influences structural assumptions. Our findings show that even despite strong structural expectations, people can overcome their initial beliefs and fully adapt to the environmental structure. Thus, our findings emphasize the dynamic and adaptable nature of structural assumptions.

7.4 Implications for causal reasoning

In the literature on causal reasoning, conditional-independence assumptions have been investigated within the causal Markov condition in Bayes nets theory (Pearl, 2000; Spirtes et al., 1993). At least two aspects of our results have implications for this literature. The first implication is that whether

people expect class-conditional independence to hold may depend on whether learning is through experience (versus read descriptions), and on whether choice behavior or explicit numerical judgments are measured. Causal reasoning studies (e.g., Mayrhofer & Waldmann, 2014; Rehder & A. B. Hoffman, 2005) often measure probability judgments about feature occurrences after giving participants an explicit description of a situation. The results from our categorization study suggest that more implicit, behavioral, and learning-based measures may reduce violations of the causal Markov condition. This finding suggests an alternative methodological approach for investigating the role of independence assumptions in causal learning and reasoning. Secondly, the dynamic adaptation of structural beliefs we found in our experiments may also hold for the degree of Markov violations in causal reasoning. Causal reasoning studies could adapt a similar paradigm by investigating causal inference in environments in which the data does or does not warrant the validity of the causal Markov condition. This approach enables systematically investigating the match between people's assumptions and inferences, the presumed causal structure of the environment, and the available learning data von Sydow, Hagmayer, and Meder, 2016, e.g.,

7.5 Beyond simplicity in early category learning

Our findings emphasize the transition between inference strategies during learning. In this sense they are consistent with the finding that in the early stages of category learning people employ a simpler inference and categorization strategy and then gradually learn more computationally intense strategies (J. D. Smith & Minda, 1998; Love et al., 2004). Our work extends these findings by adding a notion of robustness to the notion of simplicity. As outlined in the Introduction, the simple initial categorization strategy we proposed—assuming class-conditional feature independence—is additionally a robust strategy that often leads to accurate classification, despite its unrealistically simple independence assumptions (Domingos & Pazzani, 1997; Rish et al., 2001). In this sense, class-conditional independence can be viewed as a heuristic default assumption, providing an efficient means to reduce computational complexity, which works well in many situations.

Early in learning, when little information has been obtained, it is helpful to have simple strategies that facilitate making inferences and decisions in a computationally simplistic way. But simplicity is not a virtue if it only works in very few selected statistical environments. Robustness to violation of initial assumptions is also important for cognitive systems to make sound and simple inferences that safeguard against potentially costly mistakes. Simple and robust strategies for early inferences may buy time to gather more experience, and adapt to the nuances of an environment's structure. Surprisingly, the literature about strategy transitions in categorization is limited with respect to the question of whether the models proposed for early learning, for example, prototype or linearly separable models, are robust. Our results show that people may use strategies that get the best of simplicity, robustness, and adaptability.

Jarecki, & Wilke (in preparation)

Decision processes across ten evolutionary domains

Jana B. Jarecki*, Andreas Wilke†

*Max Planck Institute for Human Development, Lentzeallee 94, D-14195 Berlin
†Clarkson University, 171 Science Center, Potsdam, New York 13699-5825, USA

Abstract

The domain-specificity of human behavior is a cornerstone of evolutionary psychology, but studies investigating the actual underlying information processing are rare. The study of information processing, on the other hand, is a core content of cognitive psychology, where studies comparing different domains are rare. This project integrates these two approaches in the context of human risk-taking behavior. We measured risk attitudes using a novel evolutionary risk questionnaire and the informational cues that cause people to engage in risks using a process-tracing methodology. The domains included between-group competition, within-group competition, status/power, environmental exploration, food acquisition, food selection, kinship, parent-offspring conflict, mate attraction, and mate retention. We found that (a) domain-differences in risk-attitudes were stable, replicating earlier findings obtained from student populations now with a diverse Amazon MTurk sample; (b) the cues that respondents used for risk assessment differentiated between risk-seeking vs. risk-avoiding attitudes, and this differentiation held within domains, for the respective cues; (c) the total number of cues retrieved did not relate to the domain differences in risk attitudes; (d) the order of retrieval of risk-favoring vs. risk-avoiding cues did not relate to the domain differences in risk attitudes

1 Introduction

Many of our everyday decisions involve options with considerable variability in outcomes. Will I leave my long-term partner for a novel partner? Will I eat the marinated cockroaches at the new experimental restaurant? Will I negotiate for a higher salary with my boss during the upcoming contract prolongation meeting? Psychology has studied such choices under the umbrella of risky choice, investigating preferences for variable outcomes, framing effects, the stability of such preferences, and information integration related to them. A controversial question in the field is whether there are different decision processes in different domains of behavior or one common underlying choice process (a question considered among the 16 fundamental questions for decision research by Hastie, 2001, p. 672).

Both views, the unified and the modular, are represented in the literature. Some cognitive psychologists (Gigerenzer, Todd, & the ABC Research Group, 1999; Weber, Blais, & Betz, 2002) and many evolutionary psychologists (e.g., Cosmides & Tooby, 1992; X.-T. Wang, 1996; Barrett & Kurzban, 2006) argued for multiple algorithms or modules by which humans integrate information. The extreme of this position became known as massive modularity hypothesis. Many economists (Kahneman & Tversky, 1979; Einav et al., 2010) and other cognitive psychologists (J. R. Anderson, 1990; Busemeyer & Bruza, 2012; N. H. Anderson, 2014) proposed rather general mathematical frameworks and theories, aiming to find a unified description of judgment and decision making. This view resonates with the brain being the hardware to run cognition as all-purpose software on it.

1.1 The success of domain general theories

General cognitive theories are widespread throughout cognitive science. In categorization research, for instance, exemplar theories (Nosofsky, 1984; Medin & Schaffer, 1978) were applied to a multitude of inferences, ranging from classification of geometric shapes (Nosofsky & Clark, 1989; Nosofsky, Kruschke, & Mckinley, 1992), stylized faces (Medin & Schwanenflugel, 1981), job candidates (von Helversen & Rieskamp, 2009), wines (Scheibehenne, von Helversen, & Rieskamp, 2015), to insects (Juslin et al., 2003).

Prominent general theories of risky choice include expected value, expected subjective utility, and (cumulative) prospect theory (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992). Prospect theory describes choosing between risky options with probabilistic payoffs by a utility function with two components, one of which weights probabilities and the other evaluates outcomes relative to a reference point (formalized by weighting and value functions; Stott, 2006). This theory successfully described violations of expected value and expected utility maximization, like risk-aversion for low gains and risk-taking for high gains; and risk-aversion for high losses and risk-taking for low losses. Prospect theory was evoked to model political choices (Levy, 2003), insurance purchases, cab driver's working hour choices, investment decisions, horse-race betting, or lottery purchases (summarized in Camerer, 2004). It also predicted individual decisions over time (Glöckner & Pachur, 2012); and Trepel, Fox, and Poldrack (2005) suggested brain regions to encode prospect-theory-like functions (but

the mechanisms remain unclear, e.g., Sambrook, Roser, & Goslin, 2012).

Despite the considerable support, there are at least two major challenges for general cognitive theories. The first concerns assuming rather than testing domain generality: Studies in one domain, like investment or insurance, are of limited insight about whether the findings generalize across domains. The traditional tests of risk taking models involve only monetary scenarios. While payoff domains (gains vs. losses) are often compared with and without cognitive models, content domain comparisons are rare. Some studies find correlations of behavior across domains. Lusk and Coble (2005) presented participants with monetary lotteries and asked for their willingness to eat, buy, and accept gene-manipulated food, and found them to correlate. (Szrek, Chao, Ramlagan, & Peltzer, 2012) found only low correlations ranging from —.07 to .02 between monetary risk measures and smoking, drinking, seat-belt avoidance, and risky sexual behavior. Pachur, Hertwig, and Wolkewitz (2014) used mathematical modeling to compare parameters of prospect theory for monetary and medical losses and found more risk taking for the latter (similar, but without cognitive modeling, Y. Huang & L. Wang, 2010; Fagley & P. M. Miller, 1997). Similarly, Rettinger and Hastie (2001) compared casino, investment, legal, and university scenarios, and found differences in risk propensities. In sum, the evidence is mixed regarding whether monetary risk taking is representative for other contents. Ultimately, the degree of generality of a model across domains is an empirical question and needs explicit addressing.

The second challenge concerns content relevance. Numerous experiments present participants with risky gambles such as *win 20 points with a probability of 10%* (e.g., Glöckner & Herbold, 2011; Pachur, Hertwig, Gigerenzer, & Brandstätter, 2013; Tversky & Kahneman, 1992; Payne & Braunstein, 1978), which Lopes (1983) regarded as the pendant to the fruitfly in genetics. The feature that gambles and other risk-taking tasks like the Balloon Analogue Risk Task (e.g., Pleskac & Wershbale, 2012; Wallsten, Pleskac, & Lejuez, 2005) share is that they are abstract, numerical, and context-free. The virtue thereof is numerical precision, but less so contextual resemblance to human experience. Recently, Pleskac and Hertwig, 2014 concluded that the uniformly sampled sets of gambles in laboratory experiments are not representative of the real world, where risks and rewards tend to correlate negatively. Y. Huang and L. Wang (2010) presented their participants choices in different formats (a verbal gamble, one stressed the probabilities, another the consequences) and found that probability-focus increased risk perception of life-death risks compared to monetary risks. This challenge is shared with categorization research, which as long used de-contextualized material, like circles or triangles (e.g., Ashby & Maddox, 1992; Nosofsky, Kruschke, & Mckinley, 1992). Yet different contents yield different behavior: people differ in their categorizations of artifacts, like tools, and natural stimuli, like fruits (Verheyen, Heussen, & Storms, 2011; Z. Estes, 2003). Also, people reason more logically when contextualizing the task (Tooby & Cosmides, 1992). In conclusion, it is unclear whether the abstract formats by which research measures risk taking influences behavior. This influence may go in two directions, either people may be more consistent with abstract tasks which would artificially favor domain-general theories, or behavior may be overly heterogeneous because people consider these tasks artificial.

1.2 The Success of Domain Dependent Theories

Several authors found that the risk attitudes underlying people's decisions vary across real-life decision situations related to, for instance, the health, social, or monetary domain (Blais & Weber, 2006; Hanoch et al., 2006; Kruger, Wang, & Wilke, 2007; Wilke et al., 2014; Rettinger & Hastie, 2001; Yaniv & Foster, 1995). This work focused on behavior rather than cognitive models of information processing. In a meta analyses on the effect of cover stories on risky choice found domains to influence the effects of gain and loss framing (Kühberger, 1998). A meta-analysis by Byrnes et al. (1999) found for self-reported behavior and observed behavior that content domains determine the size of gender differences. Participants surveyed by Weber and colleagues took, on average, more risks for financial compared to ethical issues (Weber, Blais, & Betz, 2002); Rettinger and Hastie (2001) found risk averse choices with a casino cover story, risk seeking for stock investments, and risk neutrality for a university and courtroom scenario. In addition to domain differences, Wilke and collaborators (20014) showed that participants' risk taking propensities correlated with how frequently they had engaged in related behaviors in the past year or month(see also the analysis by Weber, Blais, & Betz, 2002).

This line of research faces two shortcomings as well. The first is a seemingly ad hoc choice of domains (Kühberger, 1998). The number of domains varies (four domains by Rettinger and Hastie, 2001; five by Kühberger; Byrnes et al., 1998; six by Weber, Blais, and Betz, 2002; or ten by Wilke et al., 2014), and so do their labels. How to constrain the content and number of domains? Optimally, such constraints stem from a theory largely unrelated to judgment and decision making. In this vein, Wilke et al., 2014 recently relied on evolutionary principles to individuate domains. Accordingly, the goals of the organism distinguish domains, whereby the goals are indirectly relevant to achieve inclusive fitness (which leads to functionally specified domains). The authors specified domains in two steps. They first reviewed the literature from biological anthropology and evolutionary psychology to identify evolutionarily relevant goals of risk-taking, which included domains like habitat selection, foraging, or parenting. Then they created corresponding behaviors with modern-day relevance that resembled the evolutionary goals. This results in ten domains: within-group competition, between-group competition, status/power, environmental exploration, food acquisition, food selection, parental investment, kinship, mate attraction, and mate retention. This approach is one way to address the domain-selection shortcoming.

The second shortcoming comes with methodological narrowness. Investigations of domain differences tend to be psychometric, i.e. scales measuring risk preferences or attitudes across domains. For example the DOSPERT (Weber, Blais, & Betz, 2002) is a 40-item instrument to assess risk preferences, asking about behaviors like *how likely would you be to eat 'expired' food products that still 'look okay'?* The evolutionary risk scale (Wilke et al., 2014) asks, for example, *how likely would you be to eat a piece of food that has fallen on the floor?*. Undeniably these scales are invaluable tools to establish the existence of domain differences, especially because their external validity seems high as peoples' self-reports correlate with their risk taking behavior as self-reports (e.g. Yaniv & Foster, 1995; Wilke et al., 2014) and in real-life risk-takers Hanoch et al. (2006). Yet, people's answers on a scale are of limited insight regarding the underlying processes leading to risk propensities. What follows is a lack of process tracing. While research in single domains has long tried to measure the cognitive processes of risk

taking by using, for example verbal protocols (Montgomery, 1977), eye-tracking (Su et al., 2013), cognitive capacity measures (Cokely & Kelley, 2009), or choice boards (Payne & Braunstein, 1978; Pachur, Hertwig, Gigerenzer, & Brandstätter, 2013); the attempts to trace processes across domains are scant (e.g., Rettinger & Hastie, 2001, with verbal protocols). From the questionnaire research, there is evidence that people's judgments of the benefits but not the perceived risks associate with differences in risk taking in domains (J. G. Johnson, Wilke, & Weber, 2004; Weber, Blais, & Betz, 2002). One reason for this lack of process tracing may be that the format of questionnaire items is unsuitable for methods like eye tracking; another reason may be that information processing studies attempt to present multiple variations of cues for one situation, while questionnaires present a single question per situation; and a third reason could be that it is unclear what constitutes cues in the domains.

This brief review means the problem posed above — are there multiple or one decision making process for different domains — is widely open and remains to be addressed. The present paper aims to investigate it within the evolutionary functional definitions of domains. Our goals are to (a) replicate previously found differences in risk propensities across evolutionary domains, and (b) trace the cognitive processes underlying these domain differences in risk propensities to uncover whether the cues that people retrieve from memory shed light on the genesis of domain differences in risk propensities.

To start with, we contrast several theoretical positions from the cognitive literature related to information processing in risky choice (Section 2). Then we show the stability of evolutionary risk propensities, and present the exploratory investigation of the cues that people retrieve (Section 4).

2 Linking cognitive processing to functional specification

A functional (evolutionary) view of human environments postulates that different reproductive goals relate to different types of information in different ways (Barrett & Kurzban, 2006). The goal to feed or mate might relate to signals for nutrition density or reproductive value in statistically different ways. If organisms are functional their evolved cognitive processes will then transform information differently depending on the proximate signals underlying the ultimate reproductive goals. The processing of the redness, shape, and size of an apple may differ from processing status, commitment, or attractiveness of a partner. This rests on a view of cognition first proposed by H. A. Simon (1956): mental processing matches environmental structures (see also Gigerenzer, Todd, & the ABC Research Group, 1999; Pleskac & Hertwig, 2014). A good example for the behavior to which this leads is error management theory (Haselton & Buss, 2000). Human and animals decide such that their errors minimize the larger of the potential costs (for a review, D. D. Johnson, Blumstein, Fowler, & Haselton, 2013), for example if the own energy resources are low it pays off to engage in risky status fights to avoid the comparably bigger costs of starvation.

One way in which the environment shapes decision making is through experience. The mediating cognitive process which may differ between functionally specified domains is memory storage and retrieval. In the monetary domain, studies found that experience shapes choices between gambles (reviewed by Hertwig & Erev, 2009). In these experiments participants see no information about

payoffs or probabilities but can witness a series of realized payoffs from two different gambles before they pick a gamble. When deciding from experience, people under-weight rare events. This is at variance with prospect theory (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992), which predicts over-weighting of small probabilities. From a memory perspective, which takes into account the ease of retrieval, under-weighting of rare events is expected. Among the cognitive process proposed to cause this phenomenon is contingent sampling in memory: people decide based on which previous payoffs they recall. Which information people retrieve — and also what they fail to retrieve (Schooler & Hertwig, 2005) — is incorporated in several decision process models like the recognition heuristic (Goldstein & Gigerenzer, 1999), or the fluency heuristic (Jacoby & Dallas, 1981; Schooler & Hertwig, 2005). Thus the cognitive processing variable in our study was which cues individuals retrieve from memory in ten evolutionary domains.

Generally speaking, there are two properties of the retrieved cues that could govern risk taking. The first property is the *cue direction*, that is whether a cue is positive or negative. Positive cues are cues in favor of risk taking, for example *if the floor was clean I would be more inclined to eat a fallen piece of food.*. Negative cues are cues against risk taking, for example *if the floor was dirty, I would be less likely to eat it*. Note that what differs here is only the polarity of the cue, but not the content (cleanliness of the floor).

The simplest hypothesis from a memory perspective states that in the domains where people engage in more risks, they retrieve more reasons in favor of taking the risk. Either this is because they have experienced more positive outcomes of such actions in the past; or because they have stored or retrieved positive experiences selectively in low-cost domains, but negative experiences in high-cost domains. Experiments from animals showed that how well rats learn negative associations depends on the domain of behavior (Garcia et al., 1974). In both cases, if retrieving positive or negative cues relates to domain differences, the relative frequency of cues favoring the risky action should be higher in domains with increased risk taking.

We shall not be concerned with distinguishing these two causes of retrieval, but focus in the information integration processes thereafter. Let us consider four theoretical positions regarding how the retrieved cues could be processed differently in different domains.

2.1 Tallying positive cues

The first way in which the retrieved cues could be processed is by considering the valence of the cue (whether it is positively directed towards or negatively directed against engaging in the risky behavior). These cue directions could be integrated according to a tally rule. Tally (Dawes & Corrigan, 1974) is a simple compensatory information processing heuristic (see also Gigerenzer & Goldstein, 1996) which predicts that two options are compared by counts the number of cues in favor of an option. The strategy picks the option with the highest count. Accordingly, in our scenario a tally-like cue integration predicts that two domains with different risk propensities should be distinguished by the relative number of positive to negative cues retrieved: the count of retrieved situational cues in favor of taking the risk would co-vary with the risk-taking propensities in the respective domains. Domains

with (relatively) more positive cues should be associated with higher risk propensities.

However, the converse of the above could hold as well, if people thought about the cues as justification for their behavior. Huber and Seiser (2001) found that participants use more information if explicitly asked to justify their decisions. Their own decision to state a risk preference may have set a reference point. For a person who answered a scale item with 'rather likely' there is more room in the direction of 'unlikely' than 'likely'. Attempting to justify this asymmetry, people may retrieve more hypothetical scenarios moving their choice towards 'unlikely'. Accordingly, we expect that the domains for which people retrieve more negative than positive cues are associated with relatively greater risk propensities.

2.2 Non-compensation of early cues

E. J. Johnson, Häubl, and Keinan (2007) proposed query theory, which posits that not the frequency alone but the order of retrieval matters: earlier retrieved reasons determine judgments. In a study on the endowment effect (sellers pricing goods higher than buyers) they asked people, half of which were endowed with a mug, to retrieve reasons for or against trading the mug. They found that people who recalled reasons against a transaction prior to reasons favoring it named more expensive prices than people with opposite orders (for similar order-effects in different tasks, see Weber, E. J. Johnson, Milch, C. & Goldstein, 2007; Dinner, Johnson, Goldstein, & Liu, 2011). Research with monetary gambles and decision from experience suggests that the first outcome that people recall may determine if people prefer this gamble (Madan, Ludvig, & Spetch, 2014). Also more recent experiences from the second half of the experiment predicted peoples' final choices better than the full sequence (Hertwig, Barron, et al., 2004). In terms of cue integration processes, this corresponds to a non-compensatory strategy: the order in which people consider information matters for choices (implemented in models like the recognition heuristic, Goldstein and Gigerenzer, 1999; elimination-by-aspects, Tversky, 1972; or the priority heuristic, Brandstätter et al., 2006). Following this view, we expect an order effect: the domains where people retrieve positive cues *prior to* negative cues should be associated with relatively greater risk propensities.

2.3 Most frequent cues

Lastly, a different notion holds that retrieval of qualitatively different cues relates to the domain differences in risk taking — independent of the cue direction. This idea relies on work showing that people use particular subsets of the available information that are valid predictors for a good choice, even if more information is available (implemented in the take-the-best model, Gigerenzer & Goldstein, 1999). Also animals rely on selective important pieces of evidence (for an overvies, see Hutchinson & Gigerenzer, 2005). This position is motivated by the match between cognition and environment: if in the environment a few selected cues are highly predictive of a criterion, a take-the-best strategy can be adaptively 'ecologically' rational (Todd et al., 2012) depending on environmental structures (Simsek, 2013). To predict from the environmental structure which cues should be retrieved in our scenario, one would have to conduct an analysis of all potential cues and how they relate to the magnitude of risks related to human evolution.

Table 6 summarizes the just outlined cue integration accounts.

Table 6

Alternative accounts of differential cue processing across domains

Cue Frequency	Hypothesis
Tally-Positive: relative frequency of positive cues	The greater the risk propensity in a domain, the greater the relative frequency of positive cues.
Tally-Negative: relative frequency of negative cues	The greater the risk propensity in a domain, the smaller the relative frequency of positive cues.
Lex-Direction: order of positive and negative cues	The greater the risk propensity in a domain, the earlier the retrieval of positive cues.
Lex-Quality: most frequent cues	The greater the risk propensity in a domain, the more frequent a specific cue independent of its direction.

3 Study Design

The study aimed at eliciting cues for risk, that is aspects of a situation that people find relevant for taking risks, for all ten risk domains. To this end, we measures risk attitudes and queried which situational aspects influenced people's risk taking likelihoods. A further objective of study 1 was to replicate the domain differences found in students by Wilke et al., 2014 with a more diverse sample.

3.1 Participants

One hundred and twenty six people participated, six were excluded (due to inconsistent responses), leaving a final sample of 120 (mean age 33.4 years, $SD \pm 11$, range 18 to 65 years, 62 or 52% female); they received 2 US dollars for participation (the study lasted on average 22 min, range 7 to 57 min). We recruited via Amazon Mechanical Turk; data were gathered in December 2014; the experiment was conducted according to the ethical and data protection guidelines at the Max Planck Institute for Human Development, Berlin.

3.2 Material

Risk propensities were measured with the evolutionary risk scale by Wilke et al. (2014). It measures the likelihood to engage in 30 real-life behaviors in 10 evolutionary content domains (on a seven-point likert scale from extremely unlikely to extremely likely). For example, *how likely would you be to engage in the following behavior? Eating a piece of food that has fallen on the floor*. The domains are within-group competition, between-group competition, status/power, environmental exploration, parental investment, kinship, food acquisition, food selection, mate attraction, mate retention.

The cues people retrieved were elicited using aspect-listing (E. J. Johnson, Häubl, & Keinan, 2007; Dinner et al., 2011). Aspect listing is a process-tracing technique asking people to list their reasons

for a behavior sequentially (in an open-ended format). A sample answer to the above example is: *I would be less likely if it were a food that hair could stick to easily*. To minimize justification effects, we asked participants to report conditions that would make them more or less likely to engage in the risky behaviors. We measured these conditions for only five of the 30 situations to reduce fatigue among respondents and increase data quality (Johnson, personal communication). To further reduce fatigue we dispersed the open-ended questions.[23]

Standard life-history were collected as in Wilke et al., 2014. We measured age, gender, relationship status, number of siblings, birth order, number of offspring, smallest and largest number of potential offspring, and life expectancy. These variables have been shown to systematically relate to risk propensities (X. T. Wang, Kruger, & Wilke, 2009).

3.3 Procedure

The experiment was conducted online. Participants filled out the risk scale. For the five scale items for which aspect listing was employed the procedure was as follows: Participants saw one individual questionnaire item, answered it, and were subsequently asked: *Under which conditions would you personally be more [less] likely to engage in the described behavior and under which conditions would you personally be less [more] likely to engage in the described behavior?* (the order of more/less was random). Meanwhile the screen reminded them of the original risk question text and their answer. An empty text field recorded aspects. Participants wrote one aspect at a time; pressing the enter key displayed a new empty text box; reporting was self-paced; at least one and at most ten aspects were recorded. Scale items without aspect listing were presented as a list. Participants were reminded to report both the situational aspect (e.g., *if the floor was dirty*) and the direction of change (e.g., *I would be less likely*); and instructed to list situational but not personal aspects (emotions, preferences), using an example unrelated to the subsequent risk questions. Full instructions are in A.3.1. At the end participants saw all the cues they listed for each situation and were asked to make a binary choice whether the cue increased or decreased their likelihood to engage in the risky behavior. Finally, participants reported life-history and demographic variables.

While the items of the risk scale appeared in random order, the five situations with open-ended questions were pseudo-randomly selected such that male and female respondents distributed evenly across the groups of three situations forming one domain. Each domain was rated by, on average, 27 women and 25 men (range 24 to 32 woman, 21 to 28 men). We did not offer monetary incentives for cues because this could lead to artificial reports of the maximum possible number of ten cues in each domain, but one variable of interest to us were differences in cue frequencies between domains.

3.4 Methods

To analyze the risk propensities we used mixed ordinal models, as the level of measurement of Likert scale responses is ordered multinomial data, because people treat distances between inner and outer

[23] They appeared after the answer number 1, 2, 3 and 16, 17 (of 30).

Likert scale points differently (e.g., Hamby & Levine, 2015; Lodge, 1981; Cronbach, 1946; Lantz, 2013).[24] We employed ordinal multinomial logistic regressions to analyze the effect of domains and gender on risk taking likelihood. We included gender for it is one of the most-discussed variables in the risk literature (Byrnes et al., 1999). [25] Because the risk taking likelihood was a repeated measure (each person reported it for ten domains), we used a mixed model to take the within-person response correlations into account. All models specified participants as random effects and other variables as fixed effects. We used the Akaike information criterion (AIC) to assess whether including domain increased model fit, comparing a full model ($likelihood \sim domain \times female$) to a restricted gender-only model ($likelihood \sim female$). Our analysis were conducted using the ordinal package (Christensen, 2015) in R (R Core Team, 2014).

To analyze the open-ended cues, we followed qualitative content analysis principles (Mayring, 2014) (using www.qcamap.org). This requires defining coding units (here: one statement), the analysis goal (here: which types of cues do people retrieve?), the direction of the analysis (here: the written content, rather than its effect on a reader), and sub-goals (here: obtaining types of cues, and cue directions, that is whether cues increased or decreased the risk taking likelihood). One author (JJ) generated a coding manual per domain from part of the data (see A.3.2). A research assistant helped coding. After three training sessions on 20 to 26% of responses per domain the raters coded the remaining data independently. We computed reliabilities, and then preprocessed the data, i.e. resolved coder discrepancies, and excluded erroneous statements (see A.3.2).

4 Results

We were interested in two main questions. The first concerned whether the propensity to take risks differ across domains (domain differences), and if so, whether our findings replicate those from Wilke et al. (2014) (replication). The second question related to whether the domain differences we hoped to find can be related to the frequency, order, or type of cues for risk that people retrieved (situational cues).

4.1 Domain differences

First, we compared whether the model with predictors domain and gender outperformed a model including only gender. The data were more likely under the model including both variables, controlling for the additional parameters ($AIC_{\text{full}} = 11,836$ vs. $AIC_{\text{restricted}} = 13,066$). A likelihood ratio test showed a significant log likelihood difference, $\chi^2(18) = 1266.616$, $p < .001$; thus we proceeded the analysis including domain and gender.

The obtained model coefficients revealed that self-reported risk taking propensities differed across

[24] Analysis of variance (ANOVA) has severe limits if applied to ordinal data (remarked for the binomial case already by Cochran, 1940; for summaries see Agresti, 2002; T. F. Jaeger, 2008).

[25] Intuitively, this method computes the odds of responses less or equal than the j^{th} point of the likert scale compared to responses above j for each domain ($j = 1,\ldots,6$). It then uses the ratio of two odds from two domains as a measure of whether risk taking likelihoods differs between domains (for details see Agresti, 1989).

domains and gender. The domain effects are illustrated in Figure 12 (left panel). We found main effects for all but two domains (see Table 7). Specifically we obtained higher risk taking propensities (compared to the arbitrarily chosen baseline domain between group competition) in the domains within-group competition, food selection, and kinship; and low risk taking (compared to the baseline) in the domains status/power, environmental exploration, food acquisition, and mate retention. No main effects reached significance for mate attraction and parental investment. The model further showed a negative main effect for female gender, i.e. women were less risk taking than men; but importantly several interaction effects revealed that women were *more* risk taking than men in certain domains, namely food selection, parental investment, and kinship (the women's mean interaction effect sizes in these domains compensate for the negative main effect, see Table 7).

These findings corroborates past evidence on gender and risk. In general, men show higher risk propensities than women: in a meta-analysis with 150 studies Byrnes et al. (1999) found that 60 % (of a total of 322) of the effects of gender on risk taking pointed toward men as more risk prone. Yet, the authors also revealed that the size of the effect depends on the domain. Nicholson, Soane, Fenton-O'Creevy, and Willman (2005) used the five domains of the DOSPERT and showed that while men scored higher (more risk) overall, for the domains social and career risks, women scored higher. Females were also found to exhibit higher risk preferences than men concerning life-death scenarios (Y. Huang & L. Wang, 2010). Our findings that women take more risks for food selection and parental investment behavior, while being overall less risk seeking, is in line with this work. It is noteworthy that only an analysis at the domain level can reveal the dependencies of female risk taking on the domain.

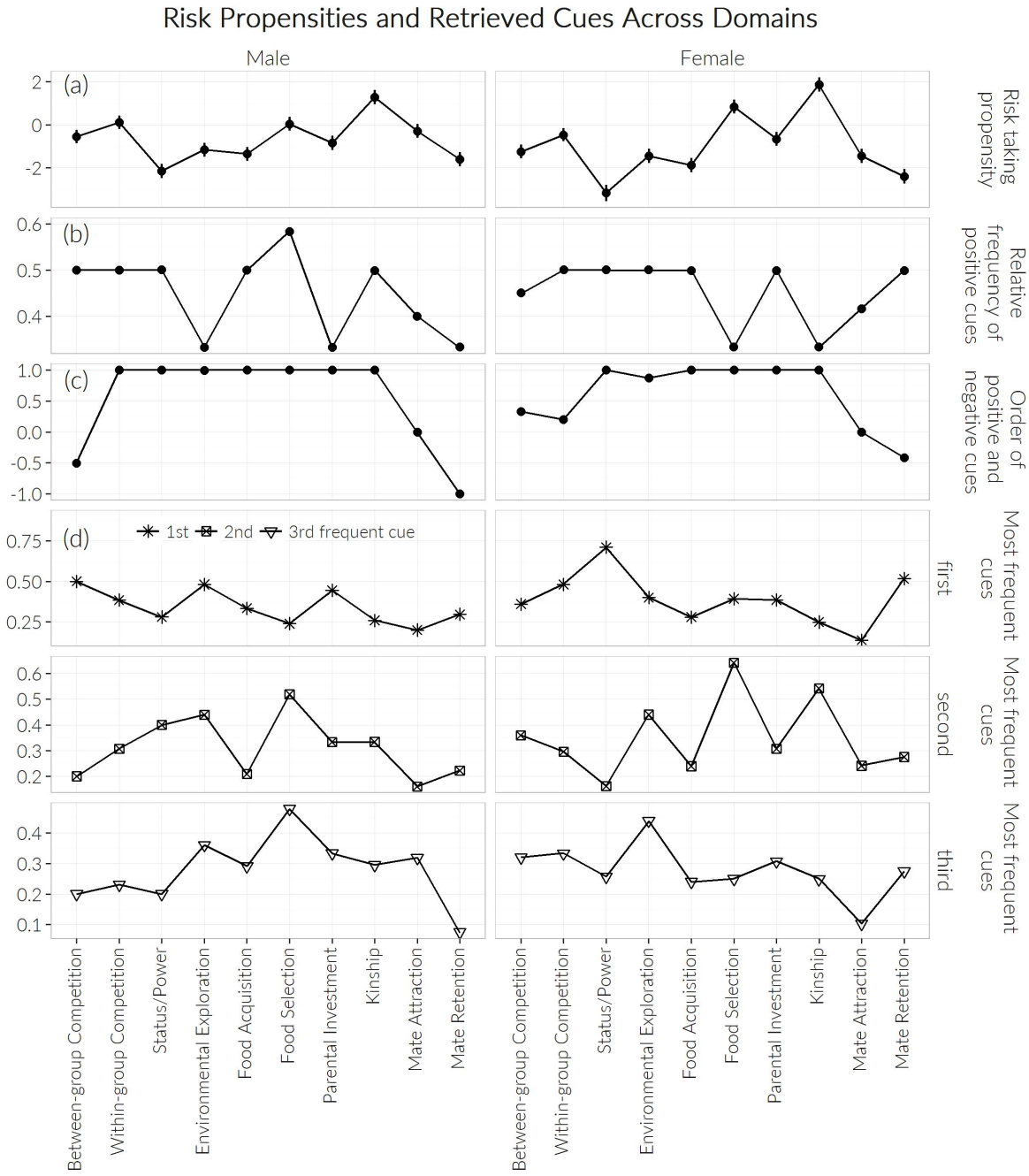

Figure 12. Likelihoods to take risks in ten domains. The figure shows the distribution of responses (likert scale from 1 "extremely unlikely" to 7 "extremely likely") split by domain. The left panel shows the data from the present study and the right from study 3 by Wilke et al (2014).

Our results are consistent with previous evidence showing that risk attitudes and perceptions differ across domains (Blais & Weber, 2006; J. G. Johnson et al., 2004; Weber, Blais, & Betz, 2002; Wilke et al., 2014). Specifically, the results that women were relatively risk taking in certain domains is theoretically predicted by evolutionary psychology, yet to the best of my knowledge date only the study by Wilke et al., 2014 explicitly tested the hypothesis.

4.2 Replication of previous studies

The next set of results investigates whether our findings replicate those from Wilke et al., 2014, obtained from multiple US colleague student populations, and investigate differences between their and our samples. The results refer to Wilke's 2014 study 3, but similar results hold using their study 2 (see A.3.3). We ran the same model (likelihood ~ domain × gender) on their data.

The results show that 8 of 10 main effects are similar between the present and the past study (see Table 7). Five of the effects of domains on risk propensities were fully replicated, namely regarding status/power, environmental exploration, food acquisition, kinship, and mate retention risks (these effects had identical signs and were both significant). In two domains, within-group competition and parental investment, the effects had identical directions (although only one reached significance); and we also replicated the negative gender main effect. In the remaining two domains, food selection and mate attraction, the present study found opposite effect directions compared to the earlier work. Our sample was less risk prone regarding mate attraction risks, that is flirtatious behavior, compared to the previous student sample. But our sample was also older and contained more married individuals than Wilke's study (see below); thus it seems unsurprising from a life-history perspective that our participants were less prone to engage in mating risks. In sum, 8 of 10 main effects had the same direction in both studies. Regarding the domain-gender interactions, all of the eight of nine effects had the same direction in both studies, and the ninth was small (and insignificant) in both studies (mate retention × gender). For seven interactions (in the domains status/power, environmental exploration, food selection, parental investment, kinship, mate attraction, and mate retention) we replicated the previous results, that is either our effects had the same direction and were significant in both studies or were small/insignificant in both studies. These replication results provide converging evidence for the dependency of human risk taking propensities on content domains.

Table 7

Replication of domain differences

	Variable	Present Study			Study 3 by Wilke et al. (2014)			Similar
		Effect	95% *CI*s	*p*	Effect	95% *CI*s	*p*	
Domain	Within-group Competition	0.659	0.302, 1.016	< .001	0.109	-0.105, 0.322	.320	Dir
	Status/Power	-1.607	-1.989, -1.226	< .001	-2.658	-2.901, -2.416	< .001	Yes
	Environmental Exploration	-0.618	-0.989, -0.247	.002	-0.359	-0.581, -0.138	.002	Yes
	Food Acquisition	-0.813	-1.187, -0.439	< .001	-0.752	-0.970, -0.534	< .001	Yes
	Food Selection	0.585	0.219, 0.952	.002	-0.280	-0.497, -0.064	.012	
	Parental Investment	-0.297	-0.667, 0.074	.117	-0.706	-0.925, -0.486	< .001	Dir
	Kinship	1.839	1.460, 2.217	< .001	1.829	1.602, 2.056	< .001	Yes
	Mate Attraction	0.258	-0.110, 0.625	.171	-1.265	-1.486, -1.043	< .001	
	Mate Retention	-1.056	-1.428, -0.684	< .001	-2.018	-2.244, -1.791	< .001	Yes
Gen.	Female	-0.700	-1.148, -0.253	.003	-0.400	-0.673, -0.126	.005	Yes
Domain × Gender	Within-group Competition × Female	0.124	-0.381, 0.630	.630	0.384	0.049, 0.719	.025	Dir
	Status/Power × Female	-0.321	-0.889, 0.246	.268	-0.171	-0.561, 0.220	.391	Yes
	Environmental Exploration × Female	0.421	-0.108, 0.950	.120	0.286	-0.062, 0.635	.108	Yes
	Food Acquisition × Female	0.181	-0.351, 0.713	.505	0.491	0.150, 0.833	.005	Dir
	Food Selection × Female	1.502	0.976, 2.028	< .001	1.180	0.840, 1.519	< .001	Yes
	Parental Investment × Female	0.886	0.355, 1.417	.002	0.746	0.398, 1.093	< .001	Yes
	Kinship × Female	1.273	0.740, 1.806	< .001	0.778	0.428, 1.127	< .001	Yes
	Mate Attraction × Female	-0.460	-0.987, 0.066	.087	-0.183	-0.533, 0.166	.304	Yes
	Mate Retention × Female	-0.097	-0.630, 0.437	.724	0.222	-0.129, 0.572	.216	Yes

Note: Comparison of the regression results (risk taking likelihood ~ domain × gender) between the present study and study 3 by Wilke et al. (2014). *Female* denotes a gender dummy with 1 = female. Effects are regression coefficients, Similar summarizes the replication: *Yes = replication* (effects have the same direction and are significant, or both are not significant at $\alpha = .05$), *Dir = direction* (effects have equal direction but only one is significant). The baseline domain (intercept) is between-group competition.

Lastly, we investigated how our sample differed from the student sample (study 3 by Wilke et al., 2014, the differences remain unchanged using study 2). Table 8 summarizes their characteristics. The participants in the present study were much older than the students (mean age 30 vs. 19 years), $t(123) = 13.987$, $p < .001$, Cohen's $d = 1.785$; the present sample had more children (mean 0.61 vs. 0.06 children per person), $t(131) = 5.428$, $p < .001$, Cohen's $d = 0.680$; and more of our participants were married (54% vs. 22% married), $\chi^2(1) = 38.576$, $p < .001$, Cohen's $h = 0.676$. Similar differences hold for the student sample in study 2.

Table 8

Demographics and life history variables in the current study and study 3 by Wilke et al., 2014

Source	Statistic	Female	Age	Married	Children	Siblings	Life Expectancy
Present sudy	Mean	51.7%	33.37	54.2%	0.61	1.56	66.98
	Range		18 – 65		0 – 5	0 – 6	35.5 – 100
Study 3 by Wilke et al.	Mean	41.6%	19.14	22.1%	0.06	1.64	81.77
	Range		18 – 45		0 – 4	0 – 5	30 – 100

4.3 Situational cues for risk taking

This sub-analysis had two goals. First, quality control of our categorization of the open-ended statements through interrater reliability measures. Second, to address whether the number of reported cues, the order of the cue direction (retrieving increasing cues first or last), or the frequency of cues was associated with risk taking likelihoods. Importantly, because not every participant reported cues for all domains, but a person reported cues for the domains 1, 2, 3, 4, 5 whereas another reported for 1, 5, 8, 9, 10; we analyzed the data on a domain-aggregated level.

4.3.1 Inter-rater reliability

Our data contained 1598 raw statements. The average interrater reliability, pooled across domains, was .69 (median .69, $SD \pm .05$, range .63 to .78), measured in terms of Gwet's AC_1 (Gwet, 2008) because we had more than two categories of unequal marginal frequencies. Reliabilities per domain are shown in A.3.4. Given that coding was rather difficult, the reliability can be considered a fair to good agreement. The final cue number after splitting and deleting errors was 1593.

4.3.2 Cue Frequency

People reported on average 2.64 cues per question (median 2.00, $SD \pm 1.29$, range 1 to 9). Table 9 shows the number of cues per question per domain, which did not differ statistically across domains, one-way ANOVA ($number \sim domain$), $F(9) = 1.893$, $p = .113$, $\eta^2 = 0.460$.

Table 10 lists the three most frequent cues per domain, and the direction in which peoples' risk taking

Table 9

Number of cues across domains, per person and situation

Domain	Mean	Median	Range
Between-group Competition	2.74	2	1, 6
Within-group Competition	2.70	2	1, 6
Status/Power	2.57	2	1, 4
Environmental Exploration	2.74	3	1, 6
Food Acquisition	2.74	2	1, 7
Food Selection	2.54	2	1, 6
Parental Investment	2.38	2	1, 5
Kinship	2.90	3	1, 9
Mate Attraction	2.27	2	1, 5
Mate Retention	2.79	3	1, 5

propensities change given these cues. The table uses only cues with aspects of the situation, excluding purely self-centered statements[26] ($N = 135$ equaling to 8.47% of all cues). Let us illustrate Table 10 with examples. In the domain mate retention people stated to be more willing to risk a romantic partnership by e.g., *not putting in the effort to fulfill the requests of your significant other* in the presence of an *external justification*: "If my phone broke","If I were involved in some kind of emergency like a snowstorm", or "If I had too much to drink". Respondents also reported increased willingness to take exploration risks, such as *going on an expedition into the desert where there will be no one else around*, if the situation provided an *incentive*: "If someone I loved was missing in the dessert", or "if the view was amazing".

Some situational cues are shared by more than one domain. *Own resources available* was retrieved in both the parental investment and the kinship domain, but points in different directions. It decreases risk taking for parental investments, e.g. asking your parents to get their old car when they get a new one, where a respondent stated to be less likely to ask "If my current car was driving fine and newer". The same type of cue increases risk taking for kinship, e.g., donating a kidney to your sibling, where a respondent stated to be more likely to donate "If my own health was good". Another cue shared by two domains, *publicity of behavior*, diminishes risk taking in both domains. Status/power, which refers to situations like blackmailing your opponent to win an election, contains this cue in the form of "If I knew that my teammate wouldn't find out". Mate attraction, which refers to situations like casually dating more than one person at a time, contains the publicity cue in the form of "If there was a chance my significant other would find out".

[26]For example "More likely if it felt good", or "Under no circumstances would I do this")

Table 10

Most frequent cues

Domain	Cue	Direction	N	Rel
Between-group Competition	Evidence for aggression	Less	13	11%
	Evidence of cooperation or friendliness	More	11	9%
	Number of peers present	More	10	8%
Within-group Competition	High stakes	More	23	14%
	Own competetive advantage	More	17	10%
	Others' competetive advantage	Less	14	8%
Status/Power	Evidence for aggression	More	31	22%
	Norm conformity of behavior	More	16	11%
	Publicity of behavior	Less	15	10%
Environmental Exploration	Number of physical obstacles	Less	28	16%
	Incentive	More	22	13%
	Number of others joining	More	22	13%
Food Acquisition	Contamination of food source	Less	18	12%
	Need for food	More	13	8%
	Familiarity with food source	More	13	8%
Food Selection	Sufficient capital	More	38	26%
	Health benefits	More	18	12%
	Price of food	Less	16	11%
Parental Investment	Usefulness of resource	More	33	25%
	Parents' resources available	More	24	18%
	Own resources available	Less	20	15%
Kinship	Own resources available	More	33	22%
	Alternative solution available	Less	16	11%
	Closeness to kin	More	12	8%
Mate Attraction	Self in relationship	Less	11	9%
	Mate wants relationship	Less	9	8%
	Publicity of behavior	Less	8	7%
Mate Retention	Damage by partner	More	28	18%
	Own commitment interest	Less	23	14%
	External justification	More	12	8%

Note: Dir = cue direction, N = Cue frequency, $Rel = N(\text{cue})/N(\text{cues in domain})$

4.3.3 Relative frequency of positive cues

Overall, about half of all cues were positive (directed towards 'more' risk taking). The median number of positive cues was 1 in 3 for women, and 1 in 2 for men, with an average percentage of 47 and 48 %, respectively; and equal medians (50%) for both genders (asymptotic Wilcoxon-Mann-Whitney Test, $Z = 0$, $p = 1$, $r = 0$). This aggregate result may be caused by a lack of motivation to report cues, however the number of cues ranges from 1 to 9, as shown in Table 9. We analyzed if individuals reported the same number of cues for all domains, and found that the range between fewest and most cues per person averaged to 1.78 cues difference, that is 1 cue in one domain and 3 in another ($t(119) = 16.417$, $p < .001$, Cohen's $d = 1.499$), which suggests that no individual-specific cue frequencies across domains.

If increased riskiness is associated with recalling more positive than negative cues, the pattern of relative frequencies of positive cues is expected to be similar to the pattern of risk taking propensities across domains. We found that positive and negative cues were about equally frequent in all domains; and the cross-domain pattern did not follow the domain specific risk taking propensities, measured as the marginal effects of domain on risk taking from the model fitted before. Figure 13 compares the pattern of risk propensities and positive cue frequencies across domains (panel (a) and (b)): The two pattern bare little resemblance, especially the relatively big willingness to take risks to help their kin had no resemblance. For women the pattern may reveal weak support for the justification-based account: it seems that the fewer positive cues the higher risk the risk propensities, yet this does not hold for many domains.

4.3.4 Order of positive and negative cues

If recalling positive prior to negative cues affects risk taking, we expect the order of cue direction and the risk taking propensities to correlate across domains. Overall, most people recalled positive before negative cues. The order of recall was assessed by the standardized mean rank difference between the positive and the negative cues (*SMRD*, computed as described in E. J. Johnson, Häubl, & Keinan, 2007):

$$\text{SMRD} = \frac{2 \cdot (\text{md}(r_p) - \text{md}(r_n))}{n}, \tag{17}$$

where md denotes the median, r_p is a vector holding the ranks of positive reasons ranked among all reasons, r_n holds the ranks of negative reasons ranked among all reasons, and n equals the number of reasons.[27] The SMRD measures the degree to which positive cues are mentioned earlier than negative ones, and takes values of -1 if all positive followed all negative cues, to $+1$, if all positive preceded all negative cues, and a value of 0 indicates a random order.

The obtained SMRD values for women and men ranged from -1 to $+1$ ($M = .21$ for both genders, medians = 1.00 and 0.76 for women and men, respectively). The difference between genders — despite their risk propensity differences — was very small (asymptotic Wilcoxon-Mann-Whitney Test, $Z =$

[27] In cases where all reasons are positive the median rank of the negative cues was set to $m_r = n + 1$; and if all reasons are negative the median rank of the positive cues was $m_r = n + 1$.

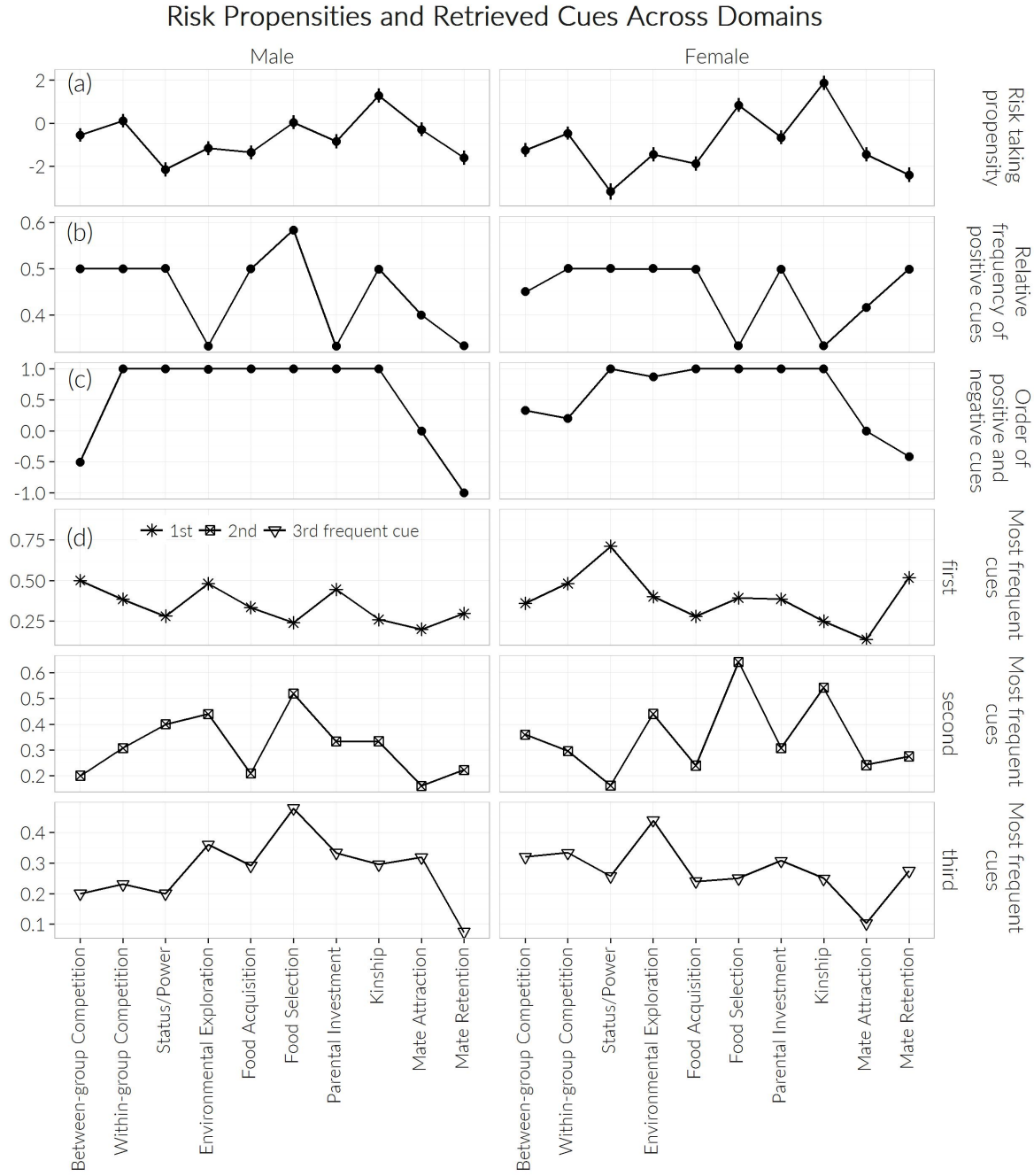

Figure 13. Risk propensities and retrieved cues across domains **(a)** Pattern of domain differences in risk propensities in terms of the mean marginal effect obtained from the regression model, error bars represent 95% confidence intervals. **(b)** Median number of positive cues per domain. **(c)** Median SMRD values per domain (see text). **(d)** Median number of individuals mentioning the first, second, and third frequent cue per domain.

0.313, $p = .754$, $r = 0.013$). The order in which people recalled positive and negative cues were similar across domains, see Figure 13 (c) shows the pattern of the orders (SMRD) across domains, which does not follow the pattern of risk taking propensities in panel (a) for neither gender.

4.3.5 Most frequent cues

Lastly, we analyzed how often the most frequent cues (see Table 10) were mentioned without regarding the cue direction. The resulting pattern across domains, shown in Figure 13 (d) reveal that for men the third-most mentioned cue shows a pattern somewhat similar to the risk taking propensities, except for the high propensity in the kinship domain (compare panel (d) to (a)). For women, the frequencies of mentioning the second-most retrieved cue tracks the domain differences in risk propensities most closely.

4.4 Discussion

To conclude, our exploration of the recalled situational aspects revealed that the differences in risk propensities across ten domains related most closely to specific cues. The frequency with which people recalled positive cues was, on average, identical for all domains. Further in high-risk domains, the number of positive cues was not greater than in low risk domains; and neither were positive cues retrieved before negative cues in those domains. While the nature of our analysis is exploratory, we believe our findings shed light on domain specific information processing.

5 General Discussion

The present study addressed the situational aspects related to human risk taking in ten evolutionary domains. It provides the following insights into human risk information processing. First, and contrary to the tenet that risk propensities are domain general, we showed that people differ systematically in their self-reported risk propensities in ten domains. Second, these differences closely replicated previous domain differences using the same ten domains although the people surveyed in our study had different demographics. A certain degree of stability in domain specific risk taking is expected from an evolutionary perspective. According to evolutionary arguments, the environment in which the cognitive system developed posed similar problems related to risk taking decisions which resulted in a set of cognitive strategies that the mind recruits still today. Stability of risk taking pattern is also expected from an ecological rationality perspective (Todd et al., 2012), if we assume that the mental processes map the structure of the environment in which people live. Third, we explored whether the situational aspects which people retrieved from memory related to the pattern of risk propensities across domains. We compared whether the cross-domain pattern of cues corresponds to the cross-domain pattern of risk propensities. Our analysis found that neither the frequency or order of cues pointing in a positive (risk-favoring) or negative (risk-avoiding) direction was associated with risk propensities across domains. Rather, specific types of cues were associated with people's domain-dependent risk propensities.

5.1 Implications

The present findings are relevant for the literature on domain-specific risk preferences. Those studies investigate domain differences with behavioral measures (Weber, Blais, & Betz, 2002; Blais & Weber, 2006). A successful approach to explain the domain differences obtained in these studies was a risk-return framework (e.g. Blais & Weber, 2006): the variables underlying risk taking behavior are the perceived costs and benefits as well as the riskiness of the situation (i.e., the variability in possible outcomes). One of the questions addressed in the literature concerned whether subjective risk perceptions or benefit perceptions are responsible for the observed domain difference in risk propensities. In some recent studies it was suggested that the process variable related to domain dependent risk propensities is the perception of expected benefits but less so the perception of the magnitude of risks (Weber, Blais, & Betz, 2002; Wilke et al., 2014). Our findings extend this evidence. The results regarding the most frequently retrieved cues show that not all cues universally refer to perceived benefits: The cues in some domains concern risk, whereas the cues in other domains concern benefits. For example, in the Between-group Competition domains the cues "evidence for aggression", and "evidence for cooperation or friendliness" are related to costs and benefits, in line with the previous findings. On the other hand, in the domain Parental Investment, the most frequent cues "Usefullness of resource" and "Own resource available" refer to benefits. This stresses the importance of process data to explain risk taking across domains and to understand at a more fine-grained level which cognitive processing variables within the risk-return framework govern risk attitudes and behavior.

Further, the fact that the present results so closely replicated past domain differences in risk attitudes, encourages the use of content domains based on biological functions. The selection of content domains in our study was based on evolutionary considerations: different domains represent modern analogues of different evolutionary adaptive challenges, which are relevant for reproductive success. This provides a theory-driven background. It would be interesting to see whether other areas of cognitive science, such as categorization, would also benefit from a standardized set of contents.

Further, our results have implications for query theory (E. J. Johnson, Häubl, & Keinan, 2007). According to this account people process information from memory by sequential queries where the result of the first queries are more strongly related to the later actions. This theory was supported with data from aspect-listing and supported for the buying and selling mugs and inter-temporal choice, (E. J. Johnson, Häubl, & Keinan, 2007; Weber, E. J. Johnson, et al., 2007). The results of the present study show that for risk propensities the order of retrieval of positive and negative cues was unrelated to the relative risk propensities across domains. This suggests two interpretations. For one, it may be that retrieval processes measured 'on line' in a task involving a choice behavior correspond to query theory more closely than the retrieval measured 'off line' with a questionnaire. Alternatively, it may be that retrieval processes regarding decisions about monetary values correspond to query theory more closely, but that decisions regarding partners, exploration, or food are driven by more than the cue direction.

One limitation of the present study concerns the validity of verbal protocols: What can self-reported lists of reasons reveal about the actual cognitive process? The evidence whether verbal data reliably

captures the actual processes is mixed. On issue is reactivity. Evidence suggests that think-aloud protocols change peoples' accuracy in selecting expected-value-maximizing risky gambles (reactivity, Russo, Johnson, & Stephens, 1989); Ericsson and H. A. Simon, 1980 review evidence against reactivity in various inference tasks; and E. J. Johnson, Häubl, and Keinan (2007) found little reactivity using aspect-listing. Thus it is advised to conduct an empirical reactivity test (Schulte-Mecklenbeck et al., 2011) by comparison to a control group. The fact that the risk propensities we found with process tracing closely replicates the results of two previous studies suggest little reactivity, which may be because aspect listing asks people for the cues retrospectively. Another issue is validity. Some argued concurrent verbalizations provide little insight (into processes of self-perception and cognitive dissonance, Nisbett & Wilson, 1977). Yet when comparing retrospective verbal reports about the information people attended to with actual eye tracking data, (Guan, Lee, Cuddihy, & Ramey, 2006) found that up to 88 % of the verbal reports matched gaze behavior. In sum, these results speak for the validity of aspect-listing as a special kind of verbal protocol with more structure than free recall.

5.2 Future directions

Our results provide several directions for future research on the processes underlying domain specific risk behavior. The first such direction concerns studies involving life-history theory (e.g. X. T. Wang et al., 2009). Life-history theory posits that individual variables such as age or number of children play a key role to determine risk taking. X. T. Wang et al. (2009) found that parenthood and age decreases risk-taking. What is unclear to date, is the question how these variables are integrated if multiple life-history variables are conflict: How would a young parent decide? Our findings that different cues matter for men and women suggests to introduce models with non-compensatory integration of life-history variables into the study of evolutionary risk-taking.

A second direction concerns model-based investigations of cue integration. The cues resulting from our process-tracing study provide a data base of empirically derived, domain specific cues for risk. Future research is needed to test whether situations involving these domains and cues lead to choices which show a pattern of risky behavior across domains that is similar to the ones obtained in the present and past studies. With this approach, cognitive models developed for multi-cue inference, such as tally (Dawes & Corrigan, 1974) or a fast-and-frugal tree (Martignon, Katsikopoulos, & Woike, 2008; Luan et al., 2011) could be tested across domains to address whether different domains come with different heuristic cue integration processes.

The third direction concerns an ecological and evolutionary validity analysis of the cues. Our findings are the first including a list of memory cues for risk taking across evolutionary domains. The cues utilized by an adaptive decision maker should reflect the structure — how cues relate to goals — of the contemporary environment at least to a degree that enables performance (Brunswik, 1955; Gigerenzer, Todd, & the ABC Research Group, 1999; Arkes et al., 2016). Whether these cues (used by a strategy) allow the decision maker to perform her goals, is an open question. This question, however, could previously not been addressed for a lack of cues. Within the evolutionary context, the decision maker's goals concern evolutionary goals, which may differ from contemporary goals and may not be directly

perceivable. Taken together, we suggest analyzing to which degree the cues we found accurately relate to the costs of taking risks in the studied domains. Additional work is necessary to address the environmental relations between the current cues and current or past costs.

Bibliography

Agresti, A. (1989). Tutorial on modeling ordered categorical response data. *Psychological Bulletin*, *105*, 290–301. doi:10.1037/0033-2909.105.2.290

Agresti, A. (2002). *Categorical Data Analysis* (3rd ed.). Hoboken, New Jersey: John Wiley & Sons.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactionson Automatic Control*, *19*, 716–723. doi:10.1109/TAC.1974.1100705

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.

Anderson, J. R. (1991a). Is human cognition adaptive? *Behavioral and Brain Sciences*, *14*, 471–485. doi:10.1017/S0140525X00070801

Anderson, J. R. (1991b). The adaptive nature of human categorization. *Psychological Review*, *98*, 409–429. doi:10.1037/0033-295X.98.3.409

Anderson, J. R. & Matessa, M. (1990). A rational analysis of categorization. In B. W. Porter & R. J. Mooney (Eds.), *Machine Learning Proceedings 1990: Proceedings of the Seventh International Conference on Machine Learning* (pp. 76–84). Austin, TX: Morgan Kaufmann Publishers Inc.

Anderson, N. H. (2014). *Unified social cognition*. New York: Psychology Press. doi:10.4324/9780203837634

Arkes, H. R., Gigerenzer, G., & Hertwig, R. (2016). How bad is incoherence? *Decision*, *3*, 20–39. doi:10.1037/dec0000043

Ashby, F. G., Alfonso-Reese, L. A., Turken, U, & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*, 442–481. doi:10.1037/0033-295X.105.3.442

Ashby, F. G. & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 33–53. doi:10.1037/0278-7393.14.1.33

Ashby, F. G. & Maddox, W. T. (1990). Integrating information from separable psychological dimensions. *Journal of Experimental Psychology: Human Perception and Performance*, *16*, 598–612. doi:10.1037/0096-1523.16.3.598

Ashby, F. G. & Maddox, W. T. (1992). Complex decision rules in categorization: Contrasting novice and experienced performance. *Journal of Experimental Psychology: Human Perception and Performance*, *18*, 50–71. doi:10.1037/0096-1523.18.1.50

Ashby, F. G. & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, *93*, 154–179. doi:10.1037//0033-295X.93.2.154

Ayal, S. & Hochman, G. (2009). Ignorance or integration: The cognitive processes underlying choice behavior. *Journal of Behavioral Decision Making*, *22*, 455–474. doi:10.1002/bdm.642

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, *44*, 211–233. doi:10.1016/0001-6918(80)90046-3

Barrett, H. C. & Kurzban, R. (2006). Modularity in cognition: framing the debate. *Psychological Review*, *113*, 628–647. doi:10.1037/0033-295X.113.3.628

Barrington, L., Marks, T. K., Hsiao, J. H.-W., & Cottrell, G. W. (2008). NIMBLE: A kernel density model of saccade-based visual memory. *Journal of Vision*, *8*, 1–14. doi:10.1167/8.14.17

Bartlema, A., Lee, M., Wetzels, R., & Vanpaemel, W. (2014). A Bayesian hierarchical mixture approach to individual differences: Case studies in selective attention and representation in category learning. *Journal of Mathematical Psychology*, *59*, 132–150. doi:10.1016/j.jmp.2013.12.002

Bellmann, R. E. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton: Princeton University Press.

Berg, N. & Gigerenzer, G. (2010). As-if behavioral economics: Neoclassical economics in disguise? *History of Economic Ideas*, *18*, 133–166. doi:10.2139/ssrn.1677168

Bergert, F. B. & Nosofsky, R. M. (2007). A response-time approach to comparing generalized rational and take-the-best models of decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 107–29. doi:10.1037/0278-7393.33.1.107

Birch, L. L. (1999). Development of food preferences. *Annual Review of Nutrition*, *19*, 41–62. doi:10.1146/annurev.nutr.19.1.41

Bischof, N. (1987). Zur Stammesgeschichte der menschlichen Kognition. *Schweizerische Zeitschrift für Psychologie*, *46*(1/2), 77–90. Retrieved from https://core.ac.uk/download/files/454/12163763.pdf

Bischof, N. (1998). *Struktur und Bedeutung: Eine Einführung in die Systemtheorie* (2nd ed.). Göttingen: Hans Huber.

Blair, M. & Homa, D. L. (2001). Expanding the search for a linear separability constraint on category learning. *Memory & Cognition*, *29*, 1153–1164. doi:10.3758/BF03206385

Blais, A.-R. & Weber, E. U. (2006). A domain-specific risk-taking (DOSPERT) scale for adult populations. *Judgment and Decision Making*, *1*, 33–47.

Bolton, G. E. & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, *90*, 166–193. doi:10.1257/aer.90.1.166

Bourne, L. E., Healy, A. F., Kole, J. A., & Graham, S. M. (2006). Strategy shifts in classification skill acquisition: Does memory retrieval dominate rule use? *Memory & Cognition*, *34*, 903–13. doi:10.3758/BF03193436

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: Making choices without trade-offs. *Psychological Review*, *113*, 409–432. doi:10.1037/0033-295X.113.2.409

Briscoe, E. & Feldman, J. (2011). Conceptual complexity and the bias/variance tradeoff. *Cognition*, *118*, 2–16. doi:10.1016/j.cognition.2010.10.004

Brown, S. D. & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*, 153–178. doi:10.1016/j.cogpsych.2007.12.002

Brunswik, E. (1943). Organismic achievement and environmental probability. *Psychological Review*, *50*, 255–272. doi:10.1037/h0060889

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, *62*, 193–217. doi:10.1037/h0047470

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). Berkeley: University of California Press.

Busemeyer, J. R. & Bruza, P. D. (2012). *Quantum Models of Decision and Cognition*. Cambridge: Cambridge University Press.

Busemeyer, J. R. & Diederich, A. (2010). *Cognitive Modeling*. Los Angeles: SAGE Publications.

Busemeyer, J. R. & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, *100*, 432–459. doi:10.1037//0033-295X.100.3.432

Byrnes, J. P., Miller, D. C., & Shafer, W. D. (1999). Gender differences in risk taking: A meta-analysis. *Psychological Bulletin*, *125*(3), 367–383.

Camerer, C. F. (2004). Prospect theory in the wild: Evidence from the field. In C. F. Camerer, G. Loewenstein, & M. Rabin (Eds.), *Advances in Behavioral Economics* (Chap. 5, pp. 148–161). Princeton: Princeton University Press.

Chase, V. M., Hertwig, R., & Gigerenzer, G. (1998). Visions of rationality. *Trends in Cognitive Sciences*, *2*(6), 206–214. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/21227174

Chater, N. (2009). Rational and mechanistic perspectives on reinforcement learning. *Cognition*, *113*, 350–364. doi:10.1016/j.cognition.2008.06.014

Chickering, D. M. & Heckerman, D. (1997). Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. *Machine Learning*, *29*, 181–212. doi:10.1023/A:1007469629108

Christensen, R. H. B. (2015). Ordinal - regression models for ordinal data. Retrieved from http://www.cran.r-project.org/package=ordinal/

Cochran, W. G. (1940). The analysis of variance when experimental errors follow the Poisson or Binomial laws. *The Annals of Mathematical Statistics*, *11*(3), 335–347. Retrieved from http://www.jstor.org/stable/2235680

Cokely, E. T. & Kelley, C. M. (2009). Cognitive abilities and superior decision making under risk: A protocol analysis and process model evaluation. *Judgment and Decision Making*, *4*, 20–33. doi:10.1016/j.jbankfin.2009.04.001

Cosmides, L. & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. T. J. Barkow, L. Cosmides (Ed.), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture* (pp. 163–228). New York: Oxford University Press. doi:10.1098/rstb.2006.1991

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, *6*, 475–494. doi:10.1177/001316444600600405

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS ONE*, *8*, e57410. doi:10.1371/journal.pone.0057410

Dawes, R. M. & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin*, *81*, 95–106. doi:10.1037/h0037613

Dinner, I., Johnson, E. J., Goldstein, D. G., & Liu, K. (2011). Partitioning default effects: why people choose not to choose. *Journal of Experimental Psychologie: Applied*, *17*, 332–41. doi:10.1037/a0024354

Domingos, P. & Pazzani, M. J. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, *29*, 103–130. doi:10.1023/A:1007413511361

Edwards, W. (1967). Conservatism in human information processing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment Under Uncertainty* (Vol. 49, *638*, pp. 359–369). Cambridge: Wiley. doi:10.1017/CBO9780511809477.026

Einav, L., Finkelstein, A., & Cullen, M. R. (2010). *How general are risk preferences? Choices under uncertainty in different domains*, Stanford Institute for Economic Policy Research. Retrieved from http://www-siepr.stanford.edu/repec/sip/09-005.pdf

Einhorn, H. J. & Hogarth, R. M. (1981). Behavioral decision theory: Processes of judgement and choice. *Annual Review of Psychology*, *19*, 1–31. doi:10.1146/annurev.ps.32.020181.000413

Einhorn, H. J., Kleinmuntz, D. N., & Kleinmuntz, B. (1979). Linear regression and process-tracing models of judgment. *Psychological Review*, *86*, 465–485. doi:10.1037//0033-295X.86.5.465

Erickson, M. A. & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*, 107–140. doi:10.1037/0096-3445.127.2.107

Ericsson, K. A. & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*, 215–251. doi:10.1037/0033-295X.87.3.215

Estes, W. K. & Maddox, W. T. (2005). Risks of drawing inferences about cognitive processes from model fits to individual versus average performance. *Psychonomic Bulletin & Review*, *12*, 403–408. doi:10.3758/BF03193784

Estes, Z. (2003). Domain differences in the structure of artifactual and natural categories. *Memory & Cognition*, *31*, 199–214. doi:10.3758/BF03194379

Fagley, N. & Miller, P. M. (1997). Framing effects and arenas of choice: Your money or your life? *Organizational Behavior and Human Decision Processes*, *71*, 355–373. doi:10.1006/obhd.1997.2725

Fehr, E. & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 817–868. doi:10.1162/003355399556151

Fehr, E. & Schmidt, K. M. (2010). On inequity aversion: A reply to Binmore and Shaked. *Journal of Economic Behavior and Organization*, *73*, 101–108. doi:10.1016/j.jebo.2009.12.001

Feng, G. C. (2014). Mistakes and how to avoid mistakes in using intercoder reliability indices. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *1*, 1–10. doi:10.1027/1614-2241/a000086

Fischbacher, U., Hertwig, R., & Bruhin, A. (2013). How to model heterogeneity in costly punishment: Insights from responders' response times. *Journal of Behavioral Decision Making*, *476*, 462–476. doi:10.1002/bdm

Flach, P. A. & Lachiche, N. (2004). Naive Bayesian classification of structured data. *Machine Learning*, *57*, 233–269. doi:10.1023/B:MACH.0000039778.69032.ab

Fleiss, J. L. & Cuzick, J. (1979). The reliability of dichotomous judgments: Unequal numbers of judges per subject. *Applied Psychological Measurement*, *3*, 537–542. doi:10.1177/014662167900300410

Frankenhuis, W. E., Panchanathan, K., & Belsky, J. (2015). A mathematical model of the evolution of individual differences in developmental plasticity arising through parental bet-hedging. *Developmental Science*, *1738*, 1–24. doi:10.1111/desc.12309

Fried, L. S. & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 234–257. doi:10.1037/0278-7393.10.2.234

Friedman, D., Massaro, D. W., Kitzis, S. N., & Cohen, M. M. (1995). A comparison of learning models. *Journal of Mathematical Psychology*, *39*, 164–178. doi:10.1006/jmps.1995.1018

Friedman, J. H. (1997). On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, *1*, 55–77. doi:10.1023/A:1009778005914

Friedman, M. (1953). The methodology of positive economics. In *Essays In Positive Economics* (Vol. 2, pp. 3–43). Chicago: University of Chicago Press.

Garcia, J., Hankins, W. G., & Rusiniak, K. W. (1974). Behavioral regulation of the milieu interne in man and rat. *Science*, *185*, 824–831. doi:10.1126/science.185.4154.824

Geis, M. L. & Zwicky, A. M. (2011). On invited inferences. *Linguistic Inquiry*, *2*(4), 561–566. Retrieved from http://www.jstor.org/stable/4177664

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*, 1–58. doi:10.1162/neco.1992.4.1.1

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases". *European Review of Social Psychology*, *2*, 83–115. doi:10.1080/14792779143000033

Gigerenzer, G. & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*, 650–669. doi:10.1037//0033-295X.103.4.650

Gigerenzer, G. & Goldstein, D. G. (1999). Betting on one good reason: The take the best heuristic. In G. Gigerenzer, P. M. Todd, & the ABC Research Group (Eds.), *Simple Heuristics That Make Us Smart* (pp. 75–95). New York: Oxford University Press.

Gigerenzer, G., Hertwig, R., & Pachur, T. (2011). *Heuristics: The Foundations of Adaptive Behavior*. New York: Oxford University Press.

Gigerenzer, G., Hoffrage, U., & Goldstein, D. G. (2008). Fast and frugal heuristics are plausible models of cognition: reply to Dougherty, Franco-Watkins, and Thomas (2008). *Psychological review*, *115*, 230–239. doi:10.1037/0033-295X.115.1.230

Gigerenzer, G. & Sturm, T. (2011). How (far) can rationality be naturalized? *Synthese*, *187*, 243–268. doi:10.1007/s11229-011-0030-6

Gigerenzer, G., Todd, P. M., & the ABC Research Group. (1999). *Simple Heuristics That Make Us Smart*. Oxford University Press.

Glöckner, A. & Betsch, T. (2008). Modeling option and strategy choices with connectionist networks: Towards an integrative model of automatic and deliberate decision making. *Judgment and Decision Making*, *3*(3), 215–228.

Glöckner, A. & Herbold, A.-K. (2011). An eye-tracking study on information processing in risky decisions: Evidence for compensatory strategies based on automatic processes. *Journal of Behavioral Decision Making*, *24*, 71–98. doi:10.1002/bdm.684

Glöckner, A. & Pachur, T. (2012). Cognitive models of risky choice: Parameter stability and predictive accuracy of prospect theory. *Cognition*, *123*, 21–32. doi:10.1016/j.cognition.2011.12.002

Glöckner, A. & Witteman, C. (2010). Beyond dual-process models: A categorisation of processes underlying intuitive judgement and decision making. *Thinking & Reasoning*, *16*, 1–25. doi:10.1080/13546780903395748

Gluth, S., Rieskamp, J., & Buchel, C. (2012). Deciding when to decide: Time-variant sequential sampling models explain the emergence of value-based decisions in the human brain. *Journal of Neuroscience*, *32*, 10686–10698. doi:10.1523/JNEUROSCI.0727-12.2012

Gluth, S., Rieskamp, J., & Büchel, C. (2014). Neural evidence for adaptive strategy selection in value-based decision-making. *Cerebral Cortex*, *24*, 2009–2021. doi:10.1093/cercor/bht049

Goldstein, D. G. & Gigerenzer, G. (1999). The recognition heuristic: How ignorance makes us smart. In *Simple Heuristics That Make Us Smart* (pp. 37–58). New York: Oxford University Press.

Goldstein, D. G. & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, *109*, 75–90. doi:10.1037//0033-295X.109.1.75

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, *26*, 213–224. doi:10.1002/bdm.1753

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*, 108–154. doi:10.1080/03640210701802071

Gregg, L. & Simon, H. (1967). Process models and stochastic theories of simple concept formation. *Journal of Mathematical Psychology*, *4*, 246–276. doi:10.1016/0022-2496(67)90052-1

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*, 357–364. doi:10.1016/j.tics.2010.05.004

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *The Cambridge Handbook of Computational Psychology* (Chap. 3, pp. 59–100). Cambridge, UK: Cambridge University Press.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2014). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmim. *Topics in Cognitive Science*, *7*, 217–229. doi:10.1111/tops.12142

Griffiths, T. L., Vul, E., & Sanborn, A. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, *21*, 263–268. doi:10.1177/0963721412447619

Guan, Z., Lee, S., Cuddihy, E., & Ramey, J. (2006). The validity of the stimulated retrospective think-aloud method as measured by eye tracking. In R. Grinter, T. Rodden, P. Aoki, E. Cutrell, R. Jeffries, & G. Olson (Eds.), *CHI 2006: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1253–1262). Montreal, Quebec, Canada: ACM Press. doi:10.1145/1124772.1124961

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*, 29–48. doi:10.1348/000711006X126600

Hamby, T. & Levine, D. S. (2015). Response-scale formats and psychological distances between categories. *Applied Psychological Measurement*, *40*, 73–75. doi:10.1177/0146621615597961

Hannan, M. T. (1985). Problems of aggregation. In H. M. Blalock (Ed.), *Causal Models in the Social Sciences* (2nd ed., Chap. 17, pp. 403–440). Hawthorne, NY: Transaction Publishers.

Hanoch, Y., Johnson, J. G., & Wilke, A. (2006). Domain specificity in experimental measures and participant recruitment: An application to risk-taking behavior. *Psychological Science*, *17*, 300–304. doi:10.1111/j.1467-9280.2006.01702.x

Haselton, M. G., Bryant, G. A., Wilke, A., Frederick, D. A., Galperin, A., Frankenhuis, W. E., & Moore, T. (2009). Adaptive rationality: An evolutionary perspective on cognitive bias. *Social Cognition*, *27*, 733–763. doi:10.1521/soco.2009.27.5.733

Haselton, M. G. & Buss, D. M. (2000). Error management theory: A new perspective on biases in cross-sex mind reading. *Journal of Personality and Social Psychology*, *78*, 81–91. doi:10.1037//0022-3514.78.1.81

Hastie, R. (2001). Problems for judgment and decision making. *Annual Review of Psychology*, *52*, 653–683. doi:10.1146/annurev.polisci.6.121901.085601

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification*. Retrieved from http://arxiv.org/abs/1502.01852

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, *15*, 534–539. doi:10.1111/j.0956-7976.2004.00715.x

Hertwig, R. & Erev, I. (2009). The description-experience gap in risky choice. *Trends in Cognitive Sciences*, *13*, 517–523. doi:10.1016/j.tics.2009.09.004

Hintze, A., Olson, R. S., Adami, C., & Hertwig, R. (2015). Risk sensitivity as an evolutionary adaptation. *Scientific Reports*, *5*. doi:10.1038/srep08242. arXiv: arXiv:1310.6338v1

Hoffman, P. J. (1960). The paramorphic representation of clinical judgment. *Psychological Bulletin*, *57*, 116–131. doi:10.1037/h0047807

Homa, D. L., Dunbar, S., & Nohre, L. (1991). Instance frequency, categorization, and the modulating effect of experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 444–458. doi:10.1037/0278-7393.17.3.444

Huang, K., Sen, S., & Szidarovszky, F. (2012). Connections among Decision Field Theory models of cognition. *Journal of Mathematical Psychology*, *56*, 287–296. doi:10.1016/j.jmp.2012.07.005

Huang, Y. & Wang, L. (2010). Sex differences in framing effects across task domain. *Personality and Individual Differences*, *48*, 649–653. doi:10.1016/j.paid.2010.01.005

Huber, O. & Seiser, G. (2001). Accounting and convincing: The effect of two types of justification on the decision process. *Journal of Behavioral Decision Making*, *14*, 69–85. doi:10.1002/1099-0771(200101)14:1<69::AID-BDM366>3.0.CO;2-T

Hutchinson, J. M. & Gigerenzer, G. (2005). Simple heuristics and rules of thumb: Where psychologists and behavioural biologists might meet. *Behavioural Processes*, *69*, 97–124. doi:10.1016/j.beproc.2005.02.019

Jacoby, L. L. & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of experimental psychology. General*, *110*, 306–340. doi:10.1037/0096-3445.110.3.306

Jaeger, M. (2003). Probabilistic classifiers and the concepts they recognize. In T. Fawcett & N. Mishra (Eds.), *Proceedings of the 20th International Conference on Machine Learning (ICML-2003)* (pp. 266–273). Washington, DC: The AAAI Press.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446. doi:10.1016/j.jml.2007.11.007

Jarecki, J. B., Tan, J. H., & Jenny, M. A. (2016). *What is a cognitive process model? A disambiguation*, Max Planck Institute for Human Development. Berlin. doi:10.2139/ssrn.2544831

Jekel, M., Glöckner, A., Fiedler, S., & Bröder, A. (2012). The rationality of different kinds of intuitive decision processes. *Synthese*, 147–160. doi:10.1007/s11229-012-0126-7

Johansen, M. K. & Palmeri, T. J. (2002). Are there representational shifts during category learning? *Cognitive Psychology*, *45*, 482–553. doi:10.1016/S0010-0285(02)00505-4

Johnson, D. D., Blumstein, D. T., Fowler, J. H., & Haselton, M. G. (2013). The evolution of error: Error management, cognitive constraints, and adaptive decision-making biases. *Trends in Ecology & Evolution*, *28*, 474–481. doi:10.1016/j.tree.2013.05.014

Johnson, D. D. & Fowler, J. H. (2013). Complexity and simplicity in the evolution of decision-making biases. *Trends in Ecology & Evolution*, *28*, 446–447. doi:10.1016/j.tree.2013.06.003

Johnson, E. J., Häubl, G., & Keinan, A. (2007). Aspects of endowment: A query theory of value construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 461–74. doi:10.1037/0278-7393.33.3.461

Johnson, E. J., Schulte-Mecklenbeck, M., & Willemsen, M. C. (2008). Process models deserve process data: Comment on Brandstätter, Gigerenzer, and Hertwig (2006). *Psychological Review*, *115*, 263–73. doi:10.1037/0033-295X.115.1.263

Johnson, J. G., Wilke, A., & Weber, E. U. (2004). Beyond a trait view of risk taking: A domain-specific scale measuring risk perceptions, expected benefits, and perceived-risk attitudes in German-speaking populations. *Polish Psychological Bulletin*, *35*, 153–163.

Jones, M. & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *The Behavioral and Brain Sciences*, *34*, 169–188. doi:10.1017/S0140525X10003134

Juslin, P., Olsson, H., & Olsson, A.-C. (2003). Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology: General*, *132*, 133–156. doi:10.1037/0096-3445.132.1.133

Kahneman, D. & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263–291. doi:10.2307/1914185

Kitano, H. (2002). Computational systems biology. *Nature*, *420*, 206–210. doi:10.1038/nature01254

Knight, F. (1921). *Risk, Uncertainty, and Profit*. Cambridge: The Riverside Press. Retrieved from http://www.econlib.org/library/Knight/knRUP.html

Knowlton, B. J., Squire, L. R., & Gluck, M. A. (1994). Probabilistic classification learning in amnesia. *Learning & Memory*, *1*, 106–120. doi:10.1101/lm.1.2.106

Krakauer, D. C. (2006). Robustness in biological systems: A provisional taxonomy. In *Complex Systems Science in Biomedicine* (Vol. 87501, *1*, pp. 183–205). Boston, MA: Springer US. doi:10.1007/978-0-387-33532-2_6

Kruger, D. J., Wang, X. T., & Wilke, A. (2007). Towards the development of an evolutionarily valid domain-specific risk-taking scale. *Evolutionary Psychology*, *5*, 555–568. doi:10.1177/147470490700500306

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22–44. doi:10.1037/0033-295X.99.1.22

Kühberger, A. (1998). The influence of framing on risky decisions: A meta-analysis. *Organizational Behavior and Human Decision Processes*, *75*, 23–55. doi:10.1006/obhd.1998.2781

Lantz, B. (2013). Equidistance of Likert-Type scales and validation of inferential methods using experiments and simulations. *Electronic Journal of Business Research Methods*, *11*(1), 16–28.

Lea, S. E. G. & Webley, P. (2006). Money as tool, money as drug: The biological psychology of a strong incentive. *Behavioral and Brain Sciences*, *29*, 161–176. doi:10.1017/S0140525X06009046

Lee, M. D. & Cummins, T. D. R. (2004). Evidence accumulation in decision making: unifying the "take the best" and the "rational" models. *Psychonomic Bulletin & Review*, *11*, 343–352. doi:10.3758/BF03196581

Lemonnier, S., Brémond, R., & Baccino, T. (2014). Discriminating cognitive processes with eye movements in a decision-making driving task. *Journal of Eye Movement Research*, *7*(4), 1–14.

Levy, J. S. (2003). Applications of prospect theory to political science. *Synthese*, *135*, 215–241. doi:10.1023/A:1023413007698

Lewandowsky, S. & Farrell, S. (2010). *Computational Modeling in Cognition: Principles and Practice*. Thousand Oaks, CAL: SAGE Publications.

Lewis, R. L., Howes, A., & Singh, S. (2014). Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in Cognitive Science*, *6*, 279–311. doi:10.1111/tops.12086

Lieder, F. & Griffiths, T. L. (2015). When to use which heuristic: A rational solution to the strategy selection problem. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Lintott, C. J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., … Vandenberg, J. (2008). Galaxy Zoo: Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, *389*, 1179–1189. doi:10.1111/j.1365-2966.2008.13689.x

Little, D. R. & Lewandowsky, S. (2009). Better learning with more error: Probabilistic feedback increases sensitivity to correlated cues in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 1041–1061. doi:10.1037/a0015902

Little, D. R., Nosofsky, R. M., Donkin, C., & Denton, S. E. (2013). Logical rules and the classification of integral-dimension stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 801–820. doi:10.1037/a0029667

Lodge, M. (1981). Magnitude Scaling: Quantitative Measurement of Opinions.

Lopes, L. L. (1983). Some thoughts on the psychological concept of risk. *Journal of Experimental Psychology: Human Perception and Performance*, *9*, 137–144. doi:10.1037/0096-1523.9.1.137

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*, 309–32. doi:10.1037/0033-295X.111.2.309

Luan, S., Schooler, L. J., & Gigerenzer, G. (2011). A signal-detection analysis of fast-and-frugal trees. *Psychological Review*, *118*, 316–338. doi:10.1037/a0022684

Lusk, J. L. & Coble, K. H. (2005). Risk perceptions, risk preference, and acceptance of risky food. *American Journal of Agricultural Economics*, *87*, 393–405. doi:10.1111/j.1467-8276.2005.00730.x

Madan, C. R., Ludvig, E. A., & Spetch, M. L. (2014). Remembering the best and worst of times: Memories for extreme outcomes bias risky decisions. *Psychonomic Bulletin & Review*, *21*, 629–636. doi:10.3758/s13423-013-0542-9

Manning, C. D., Raghavan, P., & Schutze, H. (2009). *An Introduction to Information Retrieval* (Online). Cambridge University Press. doi:10.1109/LPT.2009.2020494

Marcus, G. F. & Davis, E. (2015). Still searching for principles: A response to Goodman et al. (2015). *Psychological Science*, 21–23. doi:10.1177/0956797614568433

Marewski, J. N. & Schooler, L. J. (2011). Cognitive niches: An ecological model of strategy selection. *Psychological Review*, *118*, 393–437. doi:10.1037/a0024143

Marr, D. (1982). General Introduction. In *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (pp. 3–7). San Francisco, CA: W. H. Freeman.

Martignon, L., Katsikopoulos, K. V., & Woike, J. K. (2008). Categorization with limited resources: A family of simple heuristics. *Journal of Mathematical Psychology*, *52*, 352–361. doi:10.1016/j.jmp.2008.04.003

Martignon, L., Vitouch, O., Takezawa, M., & Forster, M. R. (2003). Naive and yet enlightened: From natural frequencies to fast and frugal decision trees. In D. Hardman & L. Macchi (Eds.), *Thinking: Psychological Perspectives on Reasoning, Judgment, and Decision Making* (Chap. 10, pp. 189–211). Chichester, UK: John Wiley & Sons.

Mata, R., von Helversen, B., & Rieskamp, J. (2011). When easy comes hard: The development of adaptive strategy selection. *Child Development*, *82*, 687–700. doi:10.1111/j.1467-8624.2010.01535.x

Mayrhofer, R., Hagmayer, Y., & Waldmann, M. R. (2010). Agents and Causes: A Bayesian error attribution model of causal reasoning. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.

Mayrhofer, R. & Waldmann, M. R. (2014). Agents and causes: Dispositional intuitions as a guide to causal structure. *Cognitive Science*, *39*, 65–95. doi:10.1111/cogs.12132

Mayring, P. (2014). *Qualitative Content Analysis*. Klagenfurt. doi:10.4135/9781446282243

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, *14*, 348–356. doi:10.1016/j.tics.2010.06.002

McDaniel, M. A., Cahill, M. J., Robbins, M., & Wiener, C. (2014). Individual differences in learning and transfer: Stable tendencies for learning exemplars versus abstracting rules. *Journal of Experimental Psychology: General*, *143*, 668–693. doi:10.1037/a0032963

Mckinley, S. C. & Nosofsky, R. M. (1995). Investigations of exemplar and decision bound models in large, ill-defined category structures. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 128–148. doi:10.1037/0096-1523.21.1.128

McNamara, J. M. & Houston, A. I. (1992). Risk-sensitive foraging: A review of the theory. *Bulletin of Mathematical Biology*, *54*, 355–378. doi:10.1016/S0092-8240(05)80031-X

McNamara, J. M., Houston, A. I., & Collins, E. J. (2001). Optimality models in behavioral biology. *SIAM Review*, *43*, 413–466. doi:10.1137/S0036144500385263

Meder, B., Hagmayer, Y., & Waldmann, M. R. (2008). Inferring interventional predictions from observational learning data. *Psychonomic Bulletin & Review*, *15*, 75–80. doi:10.3758/PBR.15.1.75

Meder, B., Hagmayer, Y., & Waldmann, M. R. (2009). The role of learning data in causal reasoning about observations and interventions. *Memory & Cognition*, *37*, 249–264. doi:10.3758/MC.37.3.249

Meder, B., Mayrhofer, R., & Waldmann, M. R. (2014). Structure induction in diagnostic causal reasoning. *Psychological Review*, *121*, 277–301. doi:10.1037/a0035944

Meder, B. & Nelson, J. D. (2012). Information search with situation-specific reward functions. *Judgment and Decision Making*, *7*(2), 119–148. Retrieved from http://journal.sjdm.org/12/12314/jdm12314.html

Medin, D. L. & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207–238. doi:10.1037//0033-295X.85.3.207

Medin, D. L. & Schwanenflugel, P. J. (1981). Linear separability in classification learning. *Journal of Experimental Psychology: Human Learning and Memory*, *7*, 355–368. doi:10.1037//0278-7393.7.5.355

Medin, D. L. & Smith, E. E. (1981). Strategies and classification learning. *Journal of Experimental Psychology: Human Learning & Memory*, *7*, 241–253. doi:10.1037//0278-7393.7.4.241

Mehta, R. & Williams, D. A. (2002). Elemental and configural processing of novel cues in deterministic and probabilistic tasks. *Learning and Motivation*, *33*, 456–484. doi:10.1016/S0023-9690(02)00008-5

Miller, G. F. & Todd, P. M. (1998). Mate choice turns cognitive. *Trends in Cognitive Sciences*, *2*, 190–198. doi:10.1016/S1364-6613(98)01169-3

Minsky, M. & Papert, S. (1969). *Perceptrons: An Introduction to Computational Geometry*. Cambridge MA: The MIT Press.

Montgomery, H. (1977). A study of intransitive preferences using a think aloud procedure. In *Decision Making and Change in Human Affairs* (1969, pp. 347–362). Dordrecht: Springer Netherlands. doi:10.1007/978-94-010-1276-8_22

Moore, A. W. (1997). The undetdetermination/indeterminacy distinction and the analytic/synthetic distinction. *Erkenntnis*, *46*, 5–32. doi:10.1023/A:1005382611551

Movellan, J. R. & McClelland, J. L. (2001). The Morton-Massaro law of information integration: Implications for models of perception. *Psychological Review*, *108*, 113–148. doi:10.1037/0033-295X.108.1.113

Myung, I. J., Balasubramanian, V., & Pitt, M. A. (2000). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences of the United States of America*, *97*, 11170–11175. doi:10.1073/pnas.170283897

Myung, J. I. & Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological Review*, *116*, 499–518. doi:10.1037/a0016104

Myung, J. I., Pitt, M. A., & Kim, W. (2003). Model Evaluation, Testing and Selection. In K. Lamberts & R. Goldstone (Eds.), *Handbook of Cognition* (Vol. 1862, pp. 1–45). London: SAGE Publications Ltd.

Nelson, J. D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, *112*, 979–999. doi:10.1037/0033-295X.112.4.979

Nelson, J. D., McKenzie, C. R. M., Cottrell, G. W., & Sejnowski, T. J. (2010). Experience matters: Information acquisition optimizes probability gain. *Psychological Science*, *21*, 960–969. doi:10.1177/0956797610372637

Newell, A. (1963). Documentation of IPL-V. *Communications of the ACM*, *6*, 86–89. doi:10.1145/366274.366296

Nicholson, N., Soane, E., Fenton-O'Creevy, M., & Willman, P. (2005). Personality and domain-specific risk taking. *Journal of Risk Research*, *8*, 157–176. doi:10.1080/1366987032000123856

Nisbett, R. E. & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*, 231–259. doi:10.1037/0033-295X.84.3.231

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 104–114. doi:10.1037/0278-7393.10.1.104

Nosofsky, R. M. (1992). Exemplar, prototypes, and similarity rules. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From Learning Theory to Connectionist Theory: Essays in Honor of William K. Estes, Vol. 1* (Chap. 8, pp. 149–167). New Jersey: Hillsdale. Retrieved from http://psycnet.apa.org/psycinfo/1992-98026-014

Nosofsky, R. M. & Bergert, F. B. (2007). Limitations of exemplar models of multi-attribute probabilistic inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 999–1019. doi:10.1037/0278-7393.33.6.999

Nosofsky, R. M. & Clark, S. E. (1989). Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 282–304. doi:10.1037/0278-7393.15.2.282

Nosofsky, R. M., Kruschke, J. K., & Mckinley, S. C. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 211–233. doi:10.1037/0278-7393.18.2.211

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*, 53–79. doi:10.1037//0033-295X.101.1.53

Nosofsky, R. M. & Stanton, R. D. (2005). Speeded classification in a probabilistic category structure: Contrasting exemplar-retrieval, decision-boundary, and prototype models. *Journal of Experimental Psychology: Human Perception and Performance*, *31*, 608–629. doi:10.1037/0096-1523.31.3.608

Nosofsky, R. M. & Zaki, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 924–940. doi:10.1037//0278-7393.28.5.924

Orquin, J. L. & Mueller Loose, S. (2013). Attention and choice: A review on eye movements in decision making. *Acta Psychologica*, *144*, 190–206. doi:10.1016/j.actpsy.2013.06.003

Pachur, T., Hertwig, R., Gigerenzer, G., & Brandstätter, E. (2013). Testing process predictions of models of risky choice: A quantitative model comparison approach. *Frontiers in Psychology*, *4*. doi:10.3389/fpsyg.2013.00646

Pachur, T., Hertwig, R., & Wolkewitz, R. (2014). The affect gap in risky choice: Affect-rich outcomes attenuate attention to probability information. *Decision*, *1*, 64–78. doi:10.1037/dec0000006

Payne, J. W. & Braunstein, M. L. (1978). Risky choice: An examination of information acquisition behavior. *Memory & Cognition*, *6*, 554–61. doi:10.3758/BF03198244

Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. New York: Cambridge University Press.

Pike, R. (1973). Response latency models for signal detection. *Psychological Review*, *80*, 53–68. doi:10.1037/h0033871

Pirolli, P. & Card, S. (1999). Information foraging. *Psychological Review*, *106*, 643–675. doi:10.1037//0033-295X.106.4.643

Pleskac, T. J. & Hertwig, R. (2014). Ecologically rational choice and the structure of the environment. *Journal of Experimental Psychology: General*, *143*, 2000–2019. doi:10.1037/h0042769

Pleskac, T. J. & Wershbale, A. (2012). Making assessments while taking tepeated tisks: A pattern of multiple response pathways. *Journal of Experimental Psychology: General*. Journal of Experimental Psychology, *143*, 142–162. doi:10.1037/a0031106

Pohl, R. (2011). On the use of recognition in inferential decision making: An overview of the debate. *Judgment and Decision Making*, *6*(5), 423–438.

Posner, M. I. & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*, 353–363. doi:10.1037/h0028558

Pothos, E. M., Busemeyer, J. R., & Trueblood, J. S. (2013). A quantum geometric model of similarity. *Psychological Review*, *120*, 679–96. doi:10.1037/a0033142

R Core Team. (2014). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org/

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108. doi:10.1037/0033-295X.85.2.59

Ratcliff, R. & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, *111*, 333–367. doi:10.1037/0033-295X.111.2.333

Ratcliff, R. & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, *9*, 438–481. doi:10.3758/BF03196302

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review*, *106*, 261–300. doi:10.1037/0033-295X.106.2.261

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*, 382–407. doi:10.1016/0010-0285(72)90014-X

Rehder, B. (2014). Independence and dependence in human causal reasoning. *Cognitive Psychology*, *72*, 54–107. doi:10.1016/j.cogpsych.2014.02.002

Rehder, B. & Burnett, R. C. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology*, *50*, 264–314. doi:10.1016/j.cogpsych.2004.09.002

Rehder, B. & Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. *Cognitive Psychology*, *51*, 1–41. doi:10.1016/j.cogpsych.2004.11.001

Reichenbach, H. (1956). *The Direction of Time* (University). Berkeley.

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The E-Z reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences*, *26*, 445–476. doi:10.1017/S0140525X03000104

Rettinger, D. a. & Hastie, R. (2001). Content effects on decision making. *Organizational Behavior and Human Decision Processes*, *85*, 336–359. doi:10.1006/obhd.2000.2948

Rieskamp, J. & Otto, P. E. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General*, *135*, 207–236. doi:10.1037/0096-3445.135.2.207

Rish, I., Hellerstein, J., & Thathachar, J. (2001). *An analysis of data characteristics that affect naive Bayes performance*. IBM Technical Report RC21993, IBM Watson Research Center.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., … Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*, 211–252. doi:10.1007/s11263-015-0816-y

Russo, J. E., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal protocols. *Memory & Cognition*, *17*, 759–769. doi:10.3758/BF03202637

Sambrook, T. D., Roser, M., & Goslin, J. (2012). Prospect theory does not describe the feedback-related negativity value function. *Psychophysiology*, *49*, 1533–44. doi:10.1111/j.1469-8986.2012.01482.x

Sanborn, A. N. (2014). Testing Bayesian and heuristic predictions of mass judgments of colliding objects. *Frontiers in Psychology*, *5*. doi:10.3389/fpsyg.2014.00938

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*, 1144–1167. doi:10.1037/a0020511

Scheibehenne, B., von Helversen, B., & Rieskamp, J. (2015). Different strategies for evaluating consumer products: Attribute- and exemplar-based approaches compared. *Journal of Economic Psychology*, *46*, 39–50. doi:10.1016/j.joep.2014.11.006

Schooler, L. J. & Hertwig, R. (2005). How forgetting aids heuristic inference. *Psychological Review*, *112*, 610–28. doi:10.1037/0033-295X.112.3.610

Schulte-Mecklenbeck, M., Kühberger, A., & Ranyard, R. (2011). *A Handbook of Process Tracing Methods for Decision Research*. New York: Taylor & Francis.

Seger, C. A. & Cincotta, C. M. (2005). The roles of the caudate nucleus in human classification learning. *The Journal of Neuroscience*, *25*, 2941–2951. doi:10.1523/JNEUROSCI.3401-04.2005

Shafto, P., Kemp, C., Mansinghka, V., & Tenenbaum, J. B. (2011). A probabilistic model of cross-categorization. *Cognition*, *120*, 1–25. doi:10.1016/j.cognition.2011.02.010

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*, 129–138. doi:10.1037/h0042769

Simon, H. A. & Kotovsky, K. (1963). Human acquisition of concepts for sequential patterns. *Psychological Review*, *70*, 534–546. doi:10.1037/h0043901

Simsek, Ö. (2013). Linear decision rule as aspiration for simple decision heuristics. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Weinberger. (Eds.), *Advances in Neural Information Processing Systems 26* (1, pp. 2904–2912). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/4888-linear-decision-rule-as-aspiration-for-simple-decision-heuristics.pdf

Sloman, S. (2005). *Causal Models: How People Think about the World and Its Alternatives*. Oxford: Oxford University Press. doi:10.1093/acprof

Sloman, S., Lombrozo, T., & Malt, B. (2003). Ontological commitments and domain specific categorization. In *Integrating the Mind: Domain General Versus Domain Specific Processes in Higher Cognition* (pp. 105–129).

Smith, J. D. & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 69–69. doi:10.1037/h0090333

Smith, J. D. & Minda, J. P. (2002). Distinguishing prototype-based and exemplar-based processes in dot-pattern category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 800–811. doi:10.1037//0278-7393.28.4.800

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, Prediction, and Search*. Cambridge, Massachussetts: MIT Press.

Stewart, N., Brown, G. D. A., & Chater, N. (2002). Sequence effects in categorization of simple perceptual stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 3–11. doi:10.1037/0278-7393.28.1.3

Stott, H. P. (2006). Cumulative prospect theory's functional menagerie. *Journal of Risk and Uncertainty*, *32*, 101–130. doi:10.1007/s11166-006-8289-6

Su, Y., Rao, L.-L., Sun, H.-Y., Du, X.-L., Li, X., & Li, S. (2013). Is making a risky choice based on a weighting and adding process? An eye-tracking investigation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1765–1780. doi:10.1037/a0032861

Sun, R. (2008). *The Cambridge Handbook of Computational Psychology*. New York: Cambridge University Press.

Svenson, O. (1979). Process descriptions of decision making. *Organizational Behavior and Human Performance*, *23*, 86–112. doi:10.1016/0030-5073(79)90048-5

Szrek, H., Chao, L.-W., Ramlagan, S., & Peltzer, K. (2012). Predicting (un)healthy behavior: A comparison of risk-taking propensity measures. *Judgment and Decision Making*, *7*(6), 716–727.

Tenenbaum, J. B. (1999). *A Bayesian framework for concept learning* (Dissertation, MIT).

Tinbergen, N. (1963). On aims and methods of Ethology. *Zeitschrift für Tierpsychologie*, *20*, 410–433. doi:10.1111/j.1439-0310.1963.tb01161.x

Titterington, D. M., Murray, G. D., Murray, L. S., Spiegelhalter, D. J., Skene, A. M., Habbema, J. D. F., & Gelpke, G. J. (1981). Comparison of discrimination techniques applied to a complex data set of head injured patients. *Journal of the Royal Statistical Society. Series A (General)*, *144*, 145. doi:10.2307/2981918

Todd, P. M., Gigerenzer, G., & the ABC Research Group. (2012). *Ecological Rationality: Intelligence in the World*. Oxford University Press. doi:10.1093/acprof

Tolman, E. C. (1925). Purpose and cognition: the determiners of animal learning. *Psychological Review*, *32*, 285–297. doi:10.1037/h0072784

Tooby, J. & Cosmides, L. (1992). The psychological foundations of culture. In J. T. J. Barkow, L. Cosmides (Ed.), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture* (pp. 19–136). New York, NY: Oxford University Press. doi:10.4324/9781410608994

Trepel, C., Fox, C. R., & Poldrack, R. A. (2005). Prospect theory on the brain? Toward a cognitive neuroscience of decision under risk. *Cognitive Brain Research*, *23*, 34–50. doi:10.1016/j.cogbrainres.2005.01.016

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review*, *79*, 281–299. doi:10.1037/h0032955

Tversky, A. & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*, 297–323. doi:10.1007/BF00122574

Usher, M. & McClelland, J. L. (2001). On the time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*, 550–592. doi:10.1037//0033-295X.108.3.550

Verheyen, S., Heussen, D., & Storms, G. (2011). On domain differences in categorization and context variety. *Memory & Cognition*, *39*, 1290–300. doi:10.3758/s13421-011-0102-3

Vigo, R. (2013). The GIST of concepts. *Cognition*, *129*, 138–162. doi:10.1016/j.cognition.2013.05.008

von Helversen, B. & Rieskamp, J. (2009). Models of quantitative estimations: Rule-based and exemplar-based processes compared. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 867–889. doi:10.1037/a0015501

von Sydow, M., Hagmayer, Y., & Meder, B. (2016). Transitive reasoning distorts induction in causal chains. *Memory & Cognition*, *44*, 469–487. doi:10.3758/s13421-015-0568-5

von Sydow, M., Meder, B., & Hagmayer, Y. (2009). A transitivity heuristic of probabilistic causal reasoning. In Niels A Taatgen & H. van Rijn (Ed.), *Proceedings of the 31st Annual Cognitive Science Society* (Vol. 1, pp. 803–808). Austin, TX: Cognitive Science Society.

Waldmann, M. R. & Martignon, L. (1998). A Bayesian network model of causal learning. In M. A. Gernsbacher & S. J. Derry (Eds.), *Proceedings of the 20th Annual Conference of the Cognitive Science Society* (pp. 1102–1107). Mahwah, NJ: Lawrence Erlbaum Associates.

Wallsten, T. S., Pleskac, T. J., & Lejuez, C. W. (2005). Modeling behavior in a clinically diagnostic sequential risk-taking task. *Psychological Review*, *112*, 862–80. doi:10.1037/0033-295X.112.4.862

Walsh, C. & Sloman, S. (2008). Updating beliefs with causal models: Violations of screening off. In M. A. Gluck, J. R. Anderson, & S. M. Kosslyn (Eds.), *Memory and Mind: A Festschrift for Gordon H. Bower* (Chap. 21, pp. 345–357). New York: Lawrence Erlbaum Associates.

Wang, X. T., Kruger, D. J., & Wilke, A. (2009). Life history variables and risk-taking propensity. *Evolution and Human Behavior*, *30*, 77–84. doi:10.1016/j.evolhumbehav.2008.09.006

Wang, X.-T. (1996). Domain-specific rationality in human choices: Violations of utility axioms and social contexts. *Cognition*, *60*, 31–63. doi:10.1016/0010-0277(95)00700-8

Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review*, *20*, 158–177. doi:10.1037/h0074428

Wattenmaker, W. D., Dewey, G. I., Murphy, T. D., & Medin, D. L. (1986). Linear separability and concept learning: Context, relational properties, and concept naturalness. *Cognitive Psychology*, *18*, 158–194. doi:10.1016/0010-0285(86)90011-3

Weber, E. U., Blais, A.-R., & Betz, N. (2002). A domain-specific risk-attitude scale: measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making*, *15*, 263–290. doi:10.1002/bdm.414

Weber, E. U. & Johnson, E. J. (2009). Mindful judgment and decision making. *Annual review of psychology*, *60*, 53–85. doi:10.1146/annurev.psych.60.110707.163633

Weber, E. U., Johnson, E. J., Milch, K. F., C., B. J., & Goldstein, D. G. (2007). Asymmetric discounting in intertemporal choice: A qyery-theory account. *Psychological Science*, *18*, 516–524. doi:10.1111/j.1467-9280.2007.01932.x

Wilke, A., Sherman, A., Curdt, B., Mondal, S., Fitzgerald, C., & Kruger, D. J. (2014). An Evolutionary Domain-Specific Risk Scale. *Evolutionary Behavioral Sciences*, *8*, 123–141. doi:10.1037/ebs0000011

Wills, A. J. & Kruschke, J. K. (2008). Models of categorization. In R. Sun (Ed.), *The Cambridge Handbook of Computational Psychology* (Chap. 9, 2002, pp. 984–1022). New York: Cambridge University Press. doi:10.1006/jmps.2001.1379

Winkel, J., Keuken, M. C., van Maanen, L., Wagenmakers, E.-J., & Forstmann, B. U. (2014). Early evidence affects later decisions: Why evidence accumulation is required to explain response time data. *Psychonomic Bulletin & Review*, Epub ahead of print. doi:10.3758/s13423-013-0551-8

Winterhalder, B. & Smith, E. A. (2000). Analyzing adaptive strategies: Human behavioral ecology at twenty-five. *Evolutionary Anthropology: Issues, News and Reviews*, *9*, 51–72. doi:10.1002/(SICI)1520-6505(2000)9:2<51::AID-EVAN1>3.3.CO;2-Z

Yaniv, I. & Foster, D. P. (1995). Graininess of judgment under uncertainty: An accuracy-informativeness trade-off. *Journal of Experimental Psychology: General*, *124*, 424–432. doi:10.1037/0096-3445.124.4.424

Zhang, H. & Ling, C. X. (2001). Geometric properties of naive Bayes in nominal domains. In *Machine Learning: {ECML} 2001, 12th European Conference on Machine Learning* (Vol. 2167, pp. 588–599). Freiburg, Germany: Springer Berlin Heidelberg. doi:10.1007/3-540-44795-4_50

List of Figures

List of Tables

Appendices

Appendix A

Supplementary materials to the studies

A.1 Jarecki, Tan, & Jenny (submitted)

A.1.1 Description of the Model Review for the Survey

The list of 116 models used in our survey (printed in A.1.2) emerged from a systematic literature search. The search proceeded in two steps (1) we identified relevant articles that test or develop models, and (2) we extracted the models tested in these articles, and selected the most prominent judgment and decision making models.

Step 1: Identification of Articles. Specifically, we searched for important or new articles proposing a model ("important" meant articles that had been cited more than 100 times, "new" meant ones that were published in 2004 or later, respectively) in two databases: Google Scholar and ISI Web of Science. We combined the fixed search term "model of" with synonyms for "decision making." The precise search phrase reads "model of * decision" OR "model of * decisions" OR model of * choice" OR "model of * choices" OR "model of * preference" OR "model of * preferences" OR "model of * inference" OR "model of * inferences" where | denotes the Boolean OR and * can be any word. To ensure source relevance, we restricted our search to *Judgment and Decision Making, Psychological Review, Journal of Experimental Psychology (JEP): General; JEP Learning, Memory, and Cognition*, and *JEP Human Perception and Performance*. This yielded 433 results. From these, we first selected all articles testing cognitive models and excluded, for example, articles proposing measurement scales or mentioning but not testing models. Of these, the first author (JBJ) selected articles on judgment and decision making, excluding, for example, articles purely about perception. As this step was somewhat open to interpretation, we randomly selected a subset of 50 articles and checked the reliability of the categorization to fields by cross-coding them (coders were JBJ, JHT); this analysis showed a high level of agreement (Cohen's kappa = .831). JBJ and JT discussed discrepancies and JBJ amended the original coding.

This procedure identified articles from the judgment and decision-making literature which dealt with models. These articles served as the sources for the models.

Step 2: Model Extraction from the Articles. We extracted the names of all models tested in the articles and looked up the models' original sources (determined by their earliest occurrence in a peer-reviewed journal or book). Importantly, model selection was independent of labels such as "process model" or "computational model."

Contemporary decision science uses — in total — 172 individual models to explain human decision

making. To ensure that the models were still relevant, we included models that had been cited more than 388 times in total or more than an average of 6.6 times per year (388 is the 66th percentile cutoff of all citations, and 6.6 is the 33rd percentile cutoff of average citations per year). We were left with 116 models, listed below.

A.1.2 List of Models from our Survey

Table A.1.1

List of from the expert survey

Model Name (in alphabetic order)	**Source**
ACT-IF, Adaptive Control of Thought in Information Foraging	(Pirolli & Card, 1999)
Additive Difference Model	(Morrison 1962; Tversky, 1969)
Additive Trade-off Model between Informativeness and Accuracy	(Yaniv & Foster, 1995)
Additive-utility Model of Delay Discounting	(Killeen, 2009)
ALM, Associative Learning Model	(DeLosh, Busemeyer, McDaniel, 1997; Busemeyer, Byun, DeLosh, McDaniel, 1997)
Anchoring and Adjustment Model	(Kahneman & Tversky, 1982; Einhorn & Hogarth, 1985)
ASCM, Adaptive Strategy Choice Model (of algebraic strategies)	(Siegler & Shipley, 1995)
Associative Accumulation Model	(Bhatia, 2013)
Attractor Model of Visual Discrimination	(Wang, 2002)
Availability Heuristic	(Tversky & Kahneman, 1973)
Beta Delta Preference Model of Temporal Discounting	(Laibson, 1997)
Biased Encoding Model (associative storage network model of memory-based judgment)	(e.g., Hastie 1980)
Brunswik's Lens Model	(Brunswik, 1956)
BSR, Bayesian Sequential Risk Taking Model	(Wallsten, Pleskac, & Lejuez, 2005; Pleskac, 2008)
Causal Bayes Nets	(Spirtes, Glymour, & Scheines, 1993; Pearl, 2000)

Table A.1.1 – continued from previous page

Complement Model of Charitable-giving	(Bernheim, 1994)
Conditional Probability Model	(Oaksford, Chater, & Larkin, 2000)
Constructed-Choice Model	(Krantz & Kunreuther, 2007)
Constructionist Theory of Inference Generation	(Graesser, Singer, & Trabasso, 1994)
CPT, Cumulative Prospect Theory	(Tversky & Kahneman, 1992)
Delta-Rule Model	(Rescorla & Wagner, 1972)
Denrell's Experience Sampling Model	(Denrell, 2005)
Dimensional Overlap Model	(Kornblum, Hasbroucq, & Osman, 1990)
Dimensional Weight Model	(Birnbaum & Stegner, 1979; Tversky, Sattath, & Slovic, 1988)
Discrete-Slot Model of Working Memory	(Zhang & Luck, 2008, 2009)
DM, Diffusion Model	(Ratcliff, 1978)
Dual Process Model of Deductive Inference	(Verschueren, Schaeken, & d'Ydewalle, 2005)
EBA, Elimination by Aspects	(Tversky, 1972)
EBM, Frequency-sensitive Exemplar Model	(Nosofsky, 1988)
EBRW, Exemplar-based Random Walk Model	(Nosofksy & Palmeri, 1997)
EGCM-RT, Extended Generalized Context Model for Reaction Times	(Lamberts, 2000)
EGCM, Extended Generalized Context Model	(Lamberts, 1998)
EW, Equal Weighting Model	(Dawes, 1979)
Exemplar Model	(Medin & Schaffer, 1978)
Exemplar-Based Network Model	(Nosofsky, Kruschke, McKinley, 1992)
Exponential Strategy Selection Model	(Rieskamp & Otto, 2006)
Extension of the Leaky, Competing Accumulator Model	(Usher & McClelland, 2004)
FA Model, Fractional Adjustment Model	(Weber, Shafir, & Blais, 2004)
Feedforward inhibition model	(Shadlen & Newsome , 2001)
Forgetting Strategy Selection Model	(Rieskamp & Otto, 2006)
FSDT, Fuzzy Signal Detection Theory	(Hancock, Masalonis, & Parasuraman, 2000)
GCM, Generalized Context Model	(Nosofsky, 1986)

Table A.1.1 – continued from previous page

General Linear Classifier	(Medin & Schwanenflugel, 1981; Ashby & Gott, 1988)
GQC, General Quadratic Classifier	(Ashby & Gott, 1988; Ashby 1992)
gRAT, Generalized Version of a Rational Model/WADD	(Nosofsky & Bergert, 2007)
gTTB, Generalized Version of Take-the-Best	(Nosofsky & Bergert, 2007)
Herrnstein's Matching Law	(Herrnstein, 1961)
HyGene	(Thomas, Dougherty, Sprenger, & Harbison, 2008)
Hyperbolic Discounting Model	(Elster, 1979)
Imagination Strategy Selection Model	(Rieskamp & Otto, 2006)
Incongruity-biased Encoding Model	(Hastie & Kumar 1979; Hastie, 1980, 1984; Scrull, 1981)
Increasing Probability Model	(Wallsten, Pleskac, Lejuez, 2005)
Independence Model of Memory and Judgment	(Anderson & Hubert, 1963; Anderson, 1981)
Independent Race Model	(Logan & Cowan, 1984)
Integrated System Model of Attention and Decision Making	(Smith & Ratcliff, 2009)
LBA, Linear Ballistic Accumulator Model	(Brown & Heathcote, 2008)
LCA, Leaky, Competing Accumulator Model	(Usher & McClelland, 2001)
Leaky Accumulator Model with Relative Criteria	(Ratcliff & Smith, 2004)
Leaky Accumulator Model	(Ratcliff & Smith, 2004)
Lexicographic Semiorder	(Luce, 1956; Tversky, 1969)
Linear Decision-boundary Model	(Ashby & Townsend, 1986)
Linear Regression	–
LISA, Learning and Inference with Schemas and Analogies	(Hummel & Holyoak, 1996; 1997)
Least Mean Square Network Model/Configural Cue Adaptive Network Model	(Gluck & Bower, 1988)
Matching Heuristic	(Dhami & Ayton, 2001)
MDFT, Multialternative Decision Field Theory	(Roe, Busemeyer, & Townsend, 2001)

Table A.1.1 – continued from previous page

MIN, Minimalist	(Gigerenzer & Goldstein, 1966; Gigerenzer et. al, 1999)
Minimum-distance Classifier	(Ashby & Townsend, 1986)
Mixture Model of Transitive Preferences	(Regenwetter, Dana, & Davis-Stober, 2011)
MMN, Max-minus-next Diffusion Model	(Ratcliff & McKoon, 1997; McMillen & Holmes, 2006)
Mutual Inhibition Model	(Usher & McClelland, 2001)
Naive Bayes Classifier	(Czerlinski, Gigerenzer, & Goldstein, 1999)
Nonstationary Process Increasing Probability Model	(Wallsten, Pleskac, Lejuez, 2005)
OU, Ornstein-Uhlenbeck Diffusion Model	(Busemeyer & Townsend, 1993)
PCS, Parallel Constraint Satisfaction Model for Probabilistic Decision Tasks	(Glöckner & Betsch, 2008)
PH, Priority Heuristic	(Brandstätter, Gigerenzer, & Hertwig, 2006)
Preference for Sequences Model	(Loewenstein & Prelec, 1993)
Present-value Comparison Model	(Ainsly, 1992)
Pretree, Preference Tree	(Tversky & Sattath, 1979)
Priority Model	(Rieskamp, 2008)
Probabilistic Contrast Model	(Cheng & Novick, 1990)
Prototype Model	(Reed; 1972)
PT, Prospect Theory	(Kahneman & Tversky, 1979)
Quantum Judgment Model	(Busemeyer, Pothos, Franco, & Trueblood, 2011)
r-Model	(Hilbig, Erdfelder, & Pohl, 2010)
RAM, Rank Affected Multiplicative Model	(Birnbaum & Stegner, 1979; Birnbaum & McIntosh, 1996)
Random Walk Model	(Stone, 1960; Laming, 1968; Link & Heath, 1975)
Recruitment Model	(LaBerge, 1992)

Table A.1.1 – continued from previous page

RELAC, Reinforcment Learning of Cognitive Strategies	(Erev & Barron, 2005)
RH, Recognition Heuristic	(Goldstein & Gigerenzer, 1999, 2002)
Rule Competition Model	(Busemeyer, & Myung, 1992)
Rule-based Categorization Models	(Nosofsky, Clark, & Shin, 1989)
RULEX, Rule-plus-exception Model	(Nosofsky, Palmeri, & McKinley, 1994)
SAMBA, Selective Attention, Mapping, and Ballistic Accumulation Model	(Brown, Marley, Donkin, & Heathcote, 2008)
SDT, Signal Detection Theory	(Tanner & Swets, 1954; Swets, Tanner, & Birdsall, 1961)
SEMAUT, Subjective Expected Multi-Attribute Utility Model	(Savage, 1954)
SS Power Model	(Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008)
SSL, Strategy Selection Learning Theory	(Rieskamp & Otto, 2006)
Stationary Process Model	(Wallsten, Pleskac, Lejuez, 2005)
Story Mode (of juror decision making)	(Pennington & Hastie, 1986, 1988)
Structural Equation Model NA Subjective Expected Utility Model	(von Neumann & Morgenstern, 1947)
SUSTAIN, Supervised and Unsupervised Stratified Adaptive Incremental Network	(Love & Medin, 1998; Love, Medin & Gureckis, 2004)
SVM, Sequential Value Matching	(Johnson & Busemeyer, 2005)
Target Model	(Wallsten, Pleskac, & Lejuez, 2005)
TAX , Transfer of Attention Exchange Model	(Birnbaum & Stegner, 1979)
Three-stage model	(Hasbroucq & Guiard, 1991)
Tradeoff Model of Intertemporal Choice	(Scholten & Read, 2010)
TTB, Take-the-best	(Gigerenzer & Goldstein, 1996)
UCIP, Unlimited Capacity Independent Parallel Processing Model	(Townsend & Wenger, 2004)
Utility Functions	
WADD, Weighted Additive Model	(Payne, Bettmann, & Johnson, 1993; Keeney & Raiffa, 1976)
Warm Glow Model of Charitable-giving	(Andreoni, 1990)

Table A.1.1 – continued from previous page

Weighting Model	–
Wiener Diffusion Model	(Stone, 1960; Laming, 1968; Link & Heath, 1975)
Wyer and Srull's Storage Bin Model	(Wyer & Srull, 1986)

A.1.3 Survey of Scientists with Modeling Experience

A.1.3.1 Instructions

In this survey, you will be asked whether some prominent models in the field are process models. Please use your preferred definition to guide your answers.

Before proceeding, please note how we identified the models in the survey:

We did a literature search for papers that tested cognitive models in the field of judgment and decision making. We identified the prominent models as those that have been published in peer-reviewed journals in judgment and decision making (e.g., Psy Review, JEP, or JDM). We included models that have been cited more than 388 times in total or more than an average of 6.6 times per year (where 388 is the 66 percentile-cutoff of all citations, and 6.6 is the 33 percentile-cutoff of average citations per year).

If you think that we have missed out an important model, please let us know in the comments section at the end of the survey.

Figure A.1.1. Main instructions of the survey

A.1.3.2 Categorization of models

Survey respondents responded to a total of 116 models.

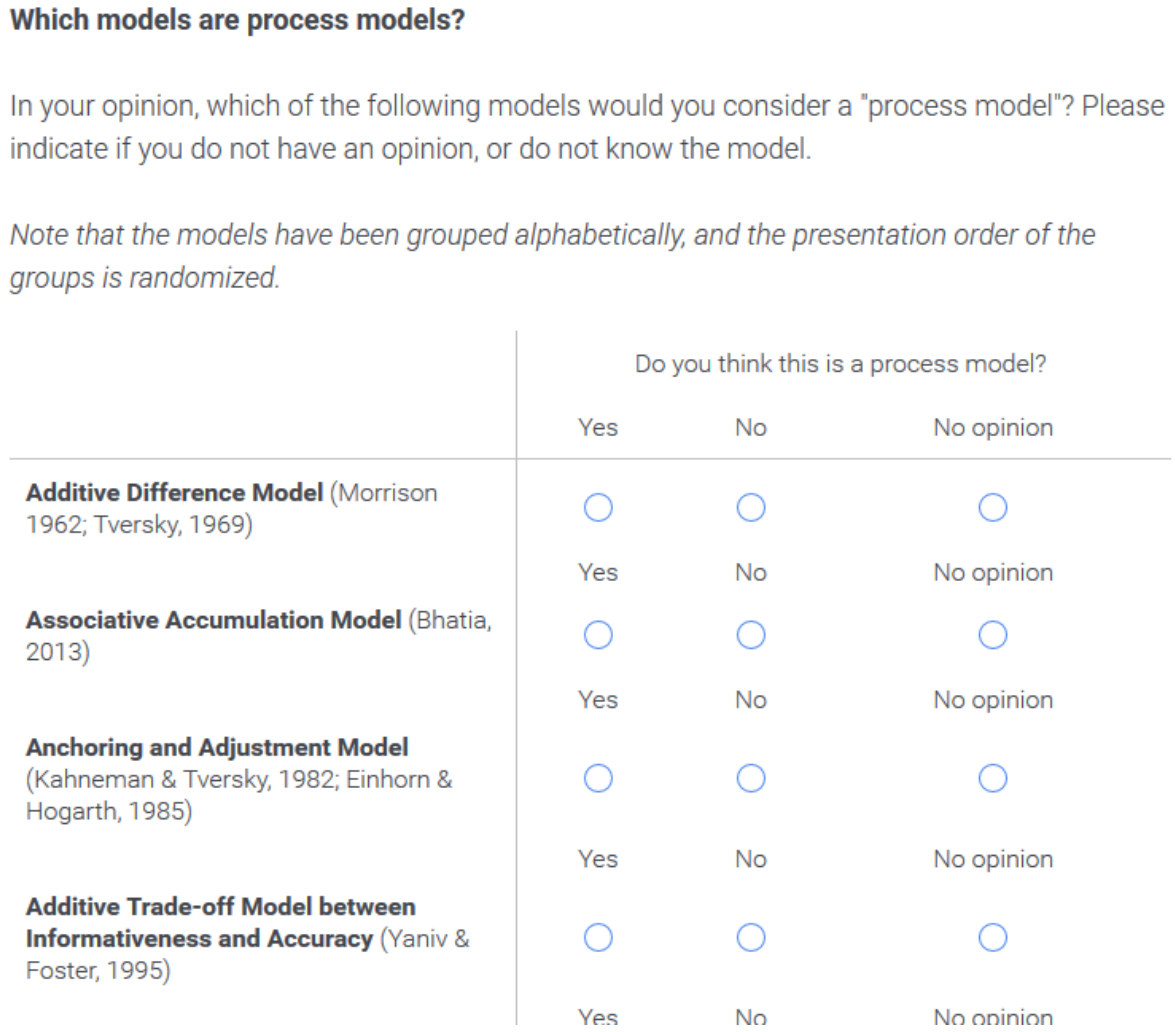

Figure A.1.2. Survey instructions: Categorization of the models

A.1.3.3 Familiarity with Marr's (1982) three levels

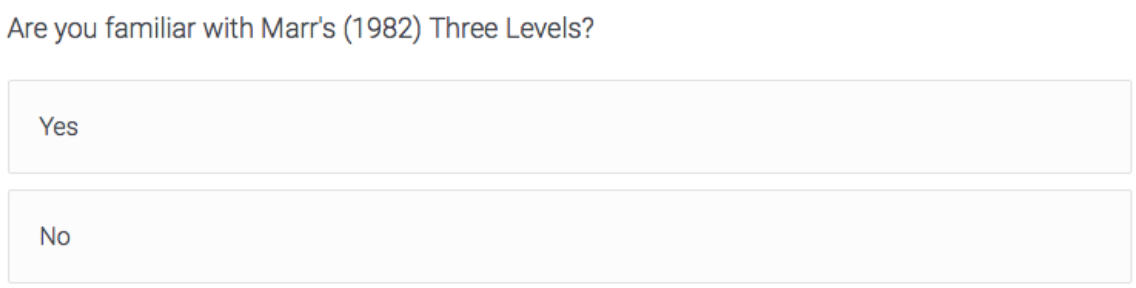

Figure A.1.3. Survey instructions: Familiarity with Marr's (1982) three levels

Only if the answer was yes to the above, the following appeared.

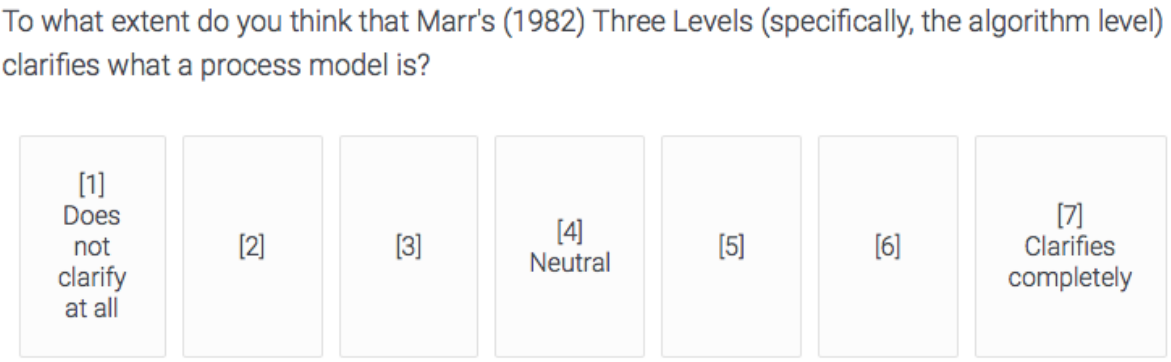

Figure A.1.4. Survey instructions: If person was familiar with Marr's (1982) three levels

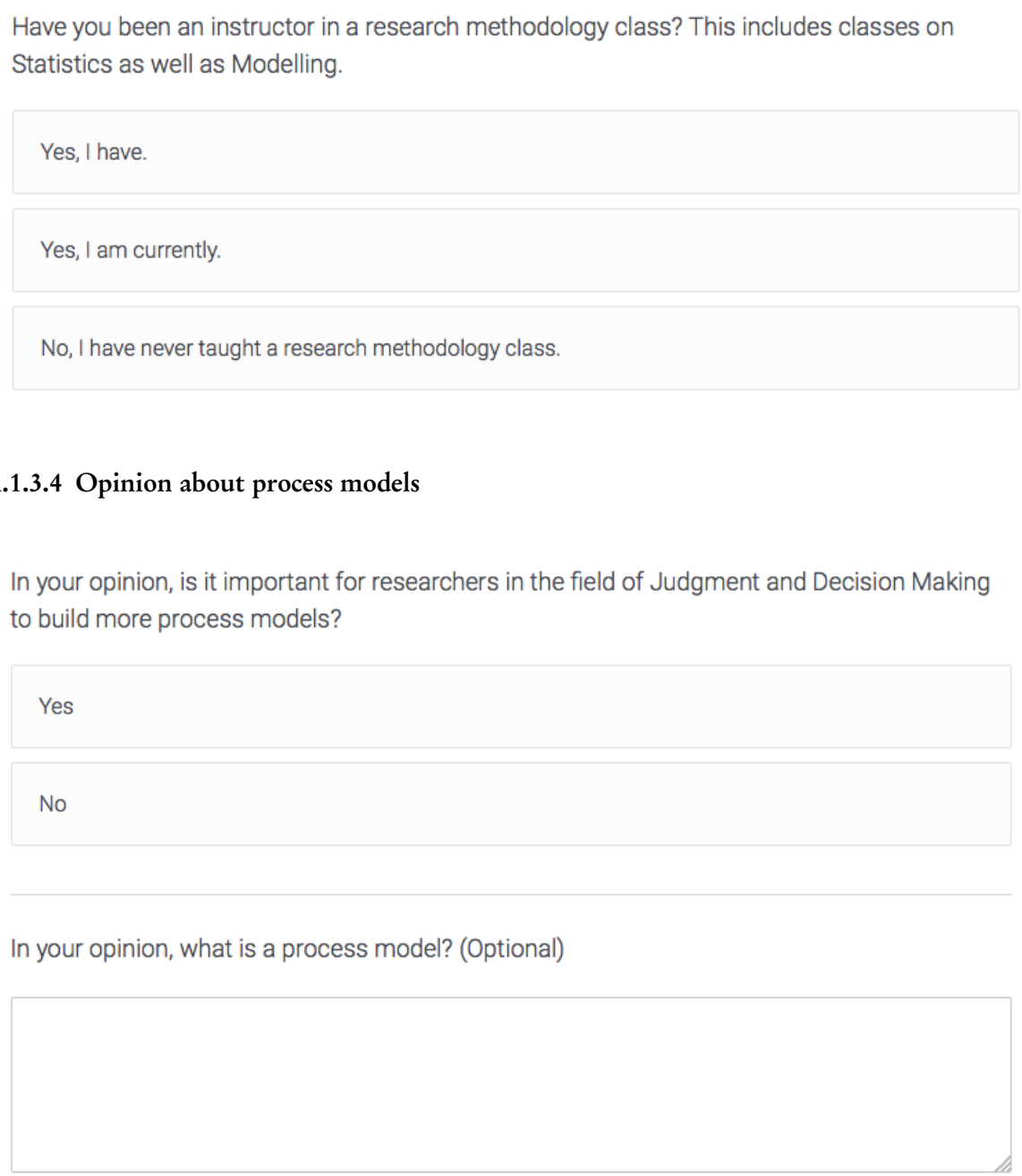

Figure A.1.5. Survey instructions: Opinion about process models

A.1.3.5 Background and Discipline

End of survey.

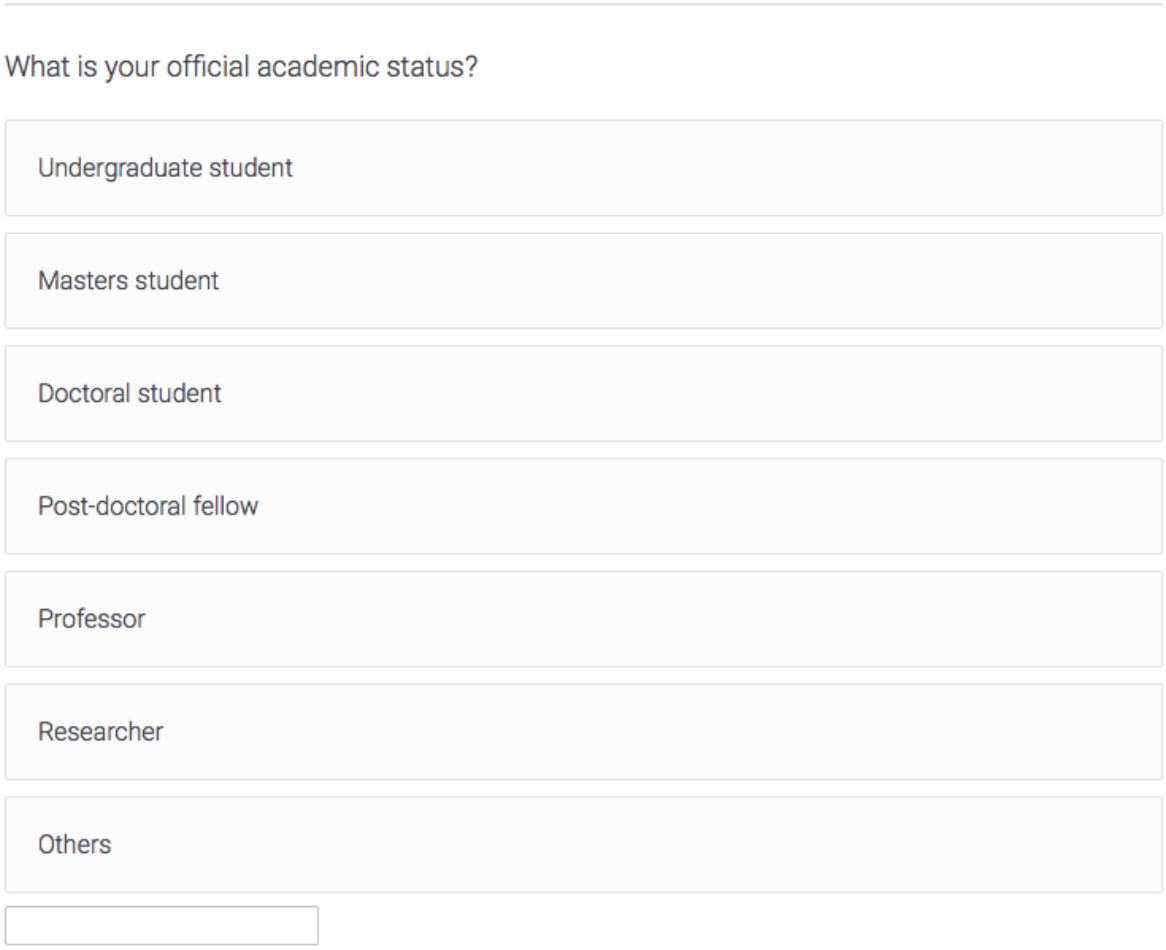

Figure A.1.6. Survey instructions: Background and discipline

A.2 Jarecki, Meder, & Nelson (accepted for publication in Cognitive Science)

A.2.1 Algorithm to find the task design

This section describes how we designed the task structure. Remember that we aimed for a task with three binary features and one binary class. We used a genetic algorithm to find parameters such that the two computations (computing the likelihoods using the configural stimuli vs. using the marginal features) entailed maximally different posterior class probabilities across the eight possible feature configurations.

The optimization was of a hill-climbing type. It selected one set of parameters randomly and computed the probability with which each stimulus belonged to class 1 in the two ways outlined above. On the basis of this result, the algorithm assigned a fitness value to the solution (see Equation A.2.1). Then it modified the set of starting parameters iteratively, aiming for higher fitness. The process repeated until convergence.

Formally, the algorithm iteratively maximized the sum of the following frequency-weighted probability difference:

$$\sum_{j=1}^{8}\Big(p(c_1 \mid s_j; cci) - p(c_1 \mid s_j; flex)\Big)^2 \cdot p(s_j)^2 \tag{A.2.1}$$

where s_j denotes the eight possible stimuli, c_1 class 1, and cci and $flex$ denote whether the posterior probability of the class was computed directly from the configural stimulus likelihoods or by multiplying the marginal feature likelihoods according to (Equation 6). The first part of the product computes the difference between the classification probability assuming class-conditional independence, $p(c_1 \mid s_j; cci)$, and the classification probability assuming flexible dependencies (i.e., arbitrary configural likelihoods), $p(c_1 \mid s_j; flex)$. The second part of the product weights the (squared) difference by the frequency of the stimulus. Squaring both terms favors large probability differences over small differences, and frequent over infrequent stimuli. Favoring rather frequent stimuli ensured that participants could actually learn the task. We set the summand to zero if both class probabilities pointed toward the same class, that is, were both $< .5$ or $> .5$.

A.2.2 Symbols

A complete list of symbols used throughout the main text and appendices is given below.

C is the class random variable, $C \in \{c_1, c_2\}$,
meaning class 1 or class 2

S is the stimulus random variable, let $S \in \{s_1, \ldots, s_8\}$
meaning stimulus 000, 001, 010, 011, 100, 101, 110, 111

$S \mid c$ is the random variable for the configural stimulus likelihoods, $S|c \in \{s_1 \mid c, \ldots, s_8 \mid c\}\ \forall\ c$
meaning stimulus 000 given A, 001 given A, 010 given A, ..., or 111 given A (when the class

$C = c_1$). When $C = c_2$ it means stimuli given class B

$F_{d|c}$ is the random variable for the marginal feature likelihoods, $F_{d|c} \in \{0, 1\} \; \forall \; d \; \forall \; c$
meaning when $d = 1$ and $C = c_1$ the first feature takes value 0 or value 1 if the stimulus belonged to class 1, the second feature takes value 0 or value 1 if the stimulus belonged to class 1 for $d = 2$ and $C = c_1$; and the third feature takes value 0 or value 1 if the stimulus belonged to class 1 if $d = 3$ and $C = c_1$. If the stimulus belonged to class 2, then $C = c_2$.

d indexes feature dimensions and $d \in \{1, 2, 3\}$
meaning the first, second, third marginal feature

t indexes trials; trials are integers starting from zero $t \in \mathbb{N}_0^+$. Trial one is $t = 0$

m indexes methods to calculate the stimulus likelihoods, and $m \in \{cci, flex\}$
cci meaning that it is calculated using class-conditional independence, and flex meaning it is calculated assuming flexible feature interactions given the class

δ is the DISC-LM's conservatism parameter and $\delta \geq 1$
meaning a free parameter in the DISC-LM[1]

π is the DISC-LM's prior belief in class-conditional feature independence, $p(M = cci)$, *before* experiencing the environment, $\pi \in [0, 1]$
It is a free parameter of the DISC-LM

w is the DISC-LM's posterior belief in class-conditional feature independence, $p(M = cci)$, *after* experiencing the environment, $w \in [0, 1]$

Notation

Greek letters denote free parameters of the DISC-LM and also the hyper-parameter in Bayesian prior distributions. Capital letters denote either random variables or counter variables to update the hyper-parameters. Small letters denote values of random variables, or indices. For simplicity and readability, we omit the capital letter random variable. We denote $p(c_1)$ to mean $P(C = c_1)$; similarly we denote $p(c_1 \mid s)$ to mean $P(C = c_1 \mid F = s)$. We use the last subscript to denote which of the eight possible stimuli a stimulus variable takes; for example, we denote s_6 to mean the sixth stimulus, or $s_6 \mid c_1$ to mean the sixth stimulus given class c. Further we use a comma to denote joint occurrences of events; that is, we denote $p(s, c)$ to mean $p(s \cap c)$; we use a semicolon to denote hierarchical dependencies as in $p(c \mid s; flex)$ denoting the probability of the class given the stimulus when the stimulus likelihood was computed assuming flexible dependencies. Capital P denotes densities; lowercase p denotes point value probabilities. Last, we distinguish a model's estimate from true values by a hat—for example, $\hat{p}(c \mid s)$ for the estimate and $p(c \mid s)$ for the true value.

[1]Note: From a mathematical standpoint, the conservatism parameter δ could be also be smaller than one. From the perspective of a model with $\delta = 1$, a model with $\delta < 1$ exhibits base-rate neglect and learns too quickly (Bar-Hillel, 1980), and a model with $\delta > 1$ shows conservatism and learns too slowly (Edwards, 1967).

A.2.3 Monte Carlo simulation procedure

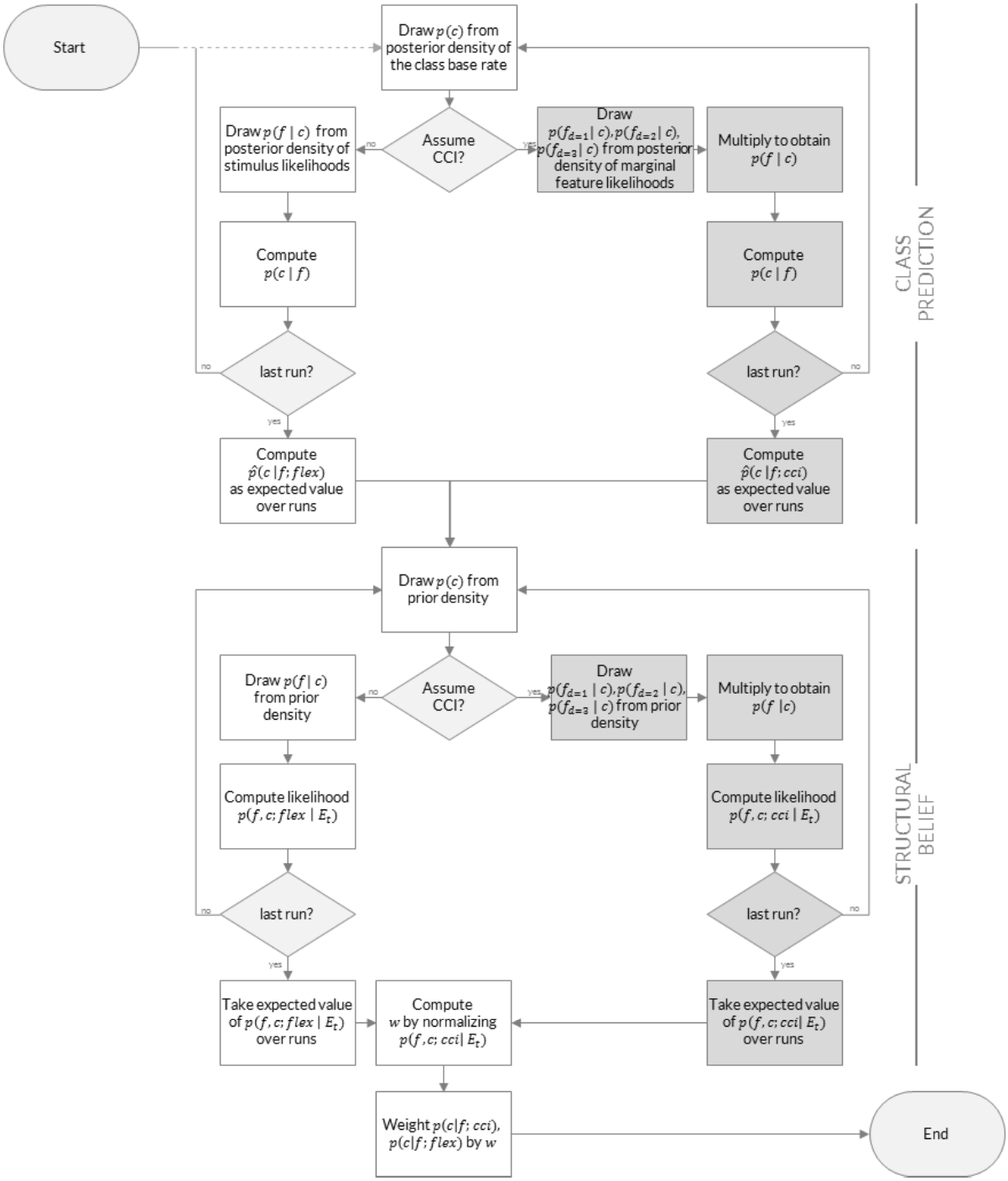

Figure A.2.1. Illustration of Monte Carlo simulation of the class prediction in the DISC-LM. *White boxes (top-left)*: How the FLEX model estimates the class prediction. *Grey boxes (top-right)*: How the CCI model estimates the class prediction. *White boxes (bottom-left)*: How the FLEX model estimates the likelihood that the world complies with its structure given the previously experienced data. *Grey boxes (bottom-right)*: How the CCI model estimates the likelihood that the world complies with its structure given the previously experienced data. In the very last row the two predictions are weighted and added in the MIX model. Note: The chart assumes that at least one stimulus and class were observed (in other words, the trial equals trial 2 or higher).

We approximated the densities with Monte Carlo simulations using 100,000 draws. The flowchart in Figure A.2.1 illustrates the steps of the Monte Carlo procedure. The upper part shows the procedure to simulate the inference of the class; the lower part shows how we simulated the match between the structural assumptions about feature independence and the data (for the Bayesian model averaging).

A.2.4 The Dependence/Independence Structure and Category-Learning Model (DISC-LM)

The model estimates the probability that the next stimulus s belongs to class 1 $p(c_1 \mid s)$. We refer to it as the class prediction. In each trial, the model computes

$$\begin{aligned} p(c_1 \mid s) &= \frac{p(s \mid c_1)\, p(c_1)}{p(s)} \\ &= \frac{p(s)\, p(c_1)}{p(s \mid c_1)p(c_1) + p(s \mid c_2)p(c_2)}, \end{aligned} \tag{A.2.2}$$

where $p(s)$ is called the stimulus likelihood of stimulus s and $p(c_1)$ the class base rate of class 1. The probability that the stimulus belongs to class 2 is $1 - p(c_1 \mid s)$.

The model infers the class base rate and the stimulus likelihoods that are required for Equation A.2.2 by Bayesian inference. In the following, we denote the stimulus likelihoods given class 1 or given class 2 as $p(s \mid c)$ and only index classes as c_1 or c_2 where necessary. We denote the class base rate of class 1 as $p(c)$.

Overview

We begin with a brief schematic overview of the conceptual differences between the inferences made by a probabilistic Bayesian learning model that takes the interactions among the features given the class into account and the inferences made by a model that assumes the features are independent, given the class (ss outlined in Section 3.1.2 and 3.1.3 in the main text). For the computational implementation details, the reader can skip the current section and proceed to the section (point estimate computation) in this appendix.

When accounting for interactions among features given the class, we infer the stimulus likelihoods, $p(s) = p(s_1 \mid c), \ldots, p(s_8 \mid c)$, given class 1 and the corresponding likelihoods given class 2 as parameters of a categorical distribution. In fact, there are seven parameters and the eighth is one minus the sum of the others. The model infers the class base rate $p(c)$ as a parameter of a Bernoulli distribution:

$$\hat{p}(c \mid s) = \frac{\overbrace{\hat{p}(s \mid c)}^{\text{Categorical}} \overbrace{\hat{p}(c)}^{\text{Bernoulli}}}{\sum_c \hat{p}(s \mid c)\hat{p}(c)}, \tag{A.2.3}$$

where $\hat{p}$ means that the probability was inferred, to distinguish the inferred from the true value p.

When assuming that features are independent given the class, we infer the class base rate as described before, but we infer the marginal feature likelihoods instead of the configural stimulus likelihoods. The feature likelihoods are multiplied to obtain the stimulus likelihoods:

$$\hat{p}(c \mid s) = \frac{\prod_d \overbrace{\hat{p}(f_d \mid c)}^{\text{Three Bernoulli}} \overbrace{\hat{p}(c)}^{\text{Bernoulli}}}{\sum_c \prod_d \hat{p}(f_d \mid c)\hat{p}(c)}, \tag{A.2.4}$$

where $d = 1, 2, 3$ indexes features, and $\hat{p}$ distinguishes the inferred from the true probability p.

Note that both methods infer the class probability, $p(c)$, as a parameter of a Bernoulli distribution, but they differ in how they infer the stimulus likelihoods.

Point estimate computation

The output of Bayesian inference is a density over probabilities, which has to be transformed into a point estimate. We do so by taking the posterior mean. We use the density of the class base rate and the density of the stimulus likelihoods, insert them into Equation A.2.2, and take the expectation. We take the posterior mean in each trial t after the density estimates are updated with the new evidence. If we denote a probability density with a capital P and a probability with a small p, the point estimate of the class prediction is computed as

$$\hat{p}_t(c \mid s) = E\left[\frac{P(\hat{p}_t(s \mid c))\, P(\hat{p}_t(c))}{\sum_c P(\hat{p}_t(s \mid c))P(\hat{p}_t(c))}\right], \tag{A.2.5}$$

where $P(\hat{p}_t(s \mid c))$ is the density of the stimulus likelihood, and $P(\hat{p}_t(c))$ is the density of the class base rate. Like above, $\hat{p}$ denotes that the parameters have been inferred.

Estimation of the posterior density of the class base rate. Let us consider how the class base rate is inferred by Bayesian inference. For an introduction to Bayesian inference, the reader is referred to Griffiths, Vul, and Sanborn, 2012. Bayesian inference integrates prior knowledge about the class in the form of a prior probability density with the knowledge gained from experiencing the classes in the environment. Let $E_t = e_1, \cdots, e_t$ be all instances of class 1 and class 2 experienced until trial t. Given E_t, the model infers the posterior density of the class base rate according to Bayes's theorem as follows:

$$P(p(c_1) \mid E_t) = \frac{P(E_t \mid p(c_1)) \cdot P(p(c_1))}{P(E_t)}. \tag{A.2.6}$$

The first term in the numerator is the density of the likelihood, the second term the prior density of the class base rate. The density in the denominator, $P(E_t)$, normalizes the product such that $0 \leq p(c_1) \leq 1$ and is computed by $P(E_t \mid p(c_1))\, P(p(c_1)) + P(E_t \mid p(c_2))\, P(p(c_2))$.

The prior density of the class 1 base rate is given by

$$P(p(c_1)) = Beta(\delta, \delta), \tag{A.2.7}$$

where $\delta \geq 1$ is a free parameter of the DISC-LM.

The posterior distribution of the class 1 base rate after experiencing the sequence of classes E_t in trial t is

$$P\left(\hat{p}_t(c_1) \mid E_t\right) = Beta(\alpha_1(t), \alpha_2(t)), \tag{A.2.8}$$

where $\alpha_1(t), \alpha_2(t)$ are shape parameters or hyper-parameters of the conjugate beta prior.

These shape parameters are given by

$$\begin{aligned} \alpha_1(t) &= \delta + \sum_t A_1(t) \\ \alpha_2(t) &= \delta + \sum_t A_2(t) \end{aligned} \tag{A.2.9}$$

where t denotes trials, δ is the conservatism parameter, and $A_1(t)$ is equal to 1 if at t the true class $c = c_1$, and 0 otherwise. Similarly, $A_2(t)$ is equal to 1 if at t the class $c = c_2$.

We performed 100,000 Monte Carlo draws from the posterior distribution $Beta(\alpha_1(t), \alpha_2(t))$ to numerically estimate the shape of the posterior density of the class base rate $\hat{p}_t(c)$ in each trial t.

Inferring the stimulus likelihoods assuming flexible dependencies of the features given the class

Let us turn to how the stimulus likelihoods are inferred when the DISC-LM makes no independence assumption (i.e., $\pi = 0$) about the interactions of the features given the class. Let us redefine $E_t = e_1, \ldots, e_t$ to be the stimuli given class 1 and class 2 that the model has experienced until trial t. Given E_t, the model infers the likelihoods $p(s \mid c) = p(s_1 \mid c), \ldots, p(s_8 \mid c)$ using Bayes's theorem by integrating the prior belief about the stimulus likelihoods $P(p(s \mid c))$ with how likely the experience is given all possible stimulus likelihoods, $P(E_t \mid p(s \mid c))$. This yields the posterior density of the stimulus likelihoods $P(p(s \mid c) \mid E_t)$:

$$P(p(s \mid c) \mid E_t) = \frac{P(E_t \mid p(s \mid c)) \cdot P(p(s \mid c))}{P(E_t)}. \tag{A.2.10}$$

The prior density of the stimulus likelihoods given class 1 is identical to the one given class 2 and is given by

$$P(p(s \mid c)) = Dirichlet(\delta, \delta, \delta, \delta, \delta, \delta, \delta, \delta). \tag{A.2.11}$$

The hyper-parameter δ is a free parameter of the DISC-LM and identical to the δ we saw before, in the prior distribution of the class base rate.

The posterior density of the stimulus likelihoods given class 1 is

$$P(\hat{p}_t(s \mid c_1)) = Dirichlet(\beta_1(t), \ldots, \beta_8(t)) \tag{A.2.12}$$

where $\beta_i(t)$ are the hyper-parameters of the Dirichlet distribution.

The hyper-parameters $\beta_i(t)$ are calculated by adding the experienced data to the prior hyper-parameter δ:

$$\begin{aligned} \beta_1(t) &= \delta + \sum_t B_1(t) \\ \vdots &= \quad \vdots \\ \beta_8(t) &= \delta + \sum_t B_8(t) \end{aligned} \tag{A.2.13}$$

where t is the last trial, B is a binary indicator variable, with $B_i(t) = 1$ if in trial t the stimulus s_i belonged to class 1, $B_i(t) = 0$ otherwise.

The posterior density of the stimulus likelihoods given class 2 is

$$P(\hat{p}_t(s \mid c_2)) = Dirichlet(\gamma_1(t), \ldots, \gamma_8(t)) \tag{A.2.14}$$

where $\gamma_i(t)$ are hyper-parameters.

The hyper-parameters $\gamma_i(t)$ are calculated by adding the experienced data to the prior hyper-parameter δ:

$$\begin{aligned} \gamma_1(t) &= \delta + \sum_t C_1(t) \\ \vdots &= \quad \vdots \\ \gamma_8(t) &= \delta + \sum_t C_8(t) \end{aligned} \tag{A.2.15}$$

where t is the last trial, C is a binary indicator variable, with $C_i(t) = 1$ if in trial t the stimulus s_i belonged to class 2, $C_i(t) = 0$ otherwise.

We performed 100,000 Monte Carlo draws from the posterior densities of the likelihoods given class 1 and given class 2, $Dirichlet(\beta_1(t),\ldots,\beta_8(t))$ and $Dirichlet(\gamma_1(t),\ldots,\gamma_8(t))$ to approximate the shape of the density of the stimulus likelihoods $\hat{p}_t(sA)$ and $\hat{p}_t(sB)$.

Inferring the stimulus likelihoods assuming independence of the features given the class

If we assume class-conditional feature independence, the stimulus probability $p(s)$ can be computed as

$$p(s)=\prod_d p(f_d), \tag{A.2.16}$$

by multiplying the *marginal* feature likelihoods $p(f_d)$ across the three stimulus dimensions $d=1,2,3$.

The marginal feature likelihoods are inferred for each feature separately. Let us define $E_{td}=e_{1d},\ldots,e_{td}$ as all values of the d^{th} feature given class c that we experienced until trial t. Given this experience, the model integrates its prior belief about the likelihood of feature d, the prior density $P(p(f_d \mid c))$, with how likely the experienced feature values are given all marginal feature likelihoods $P(E_t \mid p(f_d \mid c))$. The posterior density of the marginal feature likelihood of feature d, $P(p(f_d \mid c) \mid E_t)$, is given by

$$P(p(f_d \mid c) \mid E_t)=\frac{P(E_t \mid p(f_d \mid c)) \cdot P(p(f_d \mid c))}{p(E_t)}. \tag{A.2.17}$$

The prior distribution of the marginal feature likelihood of feature d is

$$P(p_t(f_d \mid c))=Beta(\delta,\delta) \qquad \forall\, d=1,2,3. \tag{A.2.18}$$

The inference of this model converges to the true structure of the environment if class-conditional independence actually holds.

The posterior distribution of the marginal feature likelihood of feature d given class 1 after t trials is

$$P(\hat{p}_t(f_{d,c_1}))=Beta\left(\lambda_{d,1}(t),\lambda_{d,2}(t)\right) \tag{A.2.19}$$

where d is the feature index. The parameters $\lambda_{d,1}(t)$ and $\lambda_{d,2}(t)$ are hyper-parameters.

The two hyper-parameters $\lambda_{d,1}(t)$ and $\lambda_{d,2}(t)$ are computed by adding how often a feature showed a particular value to the prior hyper-parameter δ:

$$\lambda_{d,1} = \delta + \sum_t L_{d,1}(t) \qquad \text{(A.2.20)}$$
$$\lambda_{d,2} = \delta + \sum_t L_{d,0}(t),$$

where t is the trial, δ the conservatism. The variable L is a binary indicator variable with $L_{d,1}(t) = 1$ if at trial t we observed feature d with value 1 and the stimulus belonged to class 1 in this trial, $L_{d,1}(t) = 0$ in all other cases. $L_{d,0}(t) = 1$ if in trial t we observed that feature d had value 0 and the stimulus belonged to class 1.

The posterior distribution of the marginal feature likelihood of feature d given class 2 after t trials is

$$P(\hat{p}_t(f_{d,c_2})) = Beta\left(\omega_{d,1}(t), \omega_{d,0}(t)\right)$$

where d is the feature index. The parameters $\omega_{d,1}(t)$ and $\omega_{d,2}(t)$ are hyper-parameters.

The two hyper-parameters $\omega_{d,1}(t)$ and $\omega_{d,2}(t)$ are computed by adding how often a feature showed a particular value to the prior hyper-parameter δ:

$$\omega_{d,1} = \delta + \sum_t O_{d,1}(t) \qquad \text{(A.2.21)}$$
$$\omega_{d,0} = \delta + \sum_t O_{d,2}(t)$$

where t is the trial, δ the conservatism. The variable O is a binary indicator with $O_{d,1}(t) = 1$ if at trial t we observed the d^{th} feature with value 1 and the stimulus belonged to class 2 in this trial, $O_{d,1}(t) = 0$ in all other cases. $O_{d,0}(t) = 1$ if in trial t we observed that feature d had value 0 and the stimulus belonged to class 2.

In each trial t, we performed 100,000 Monte Carlo draws from the posterior densities of the feature likelihoods, $Beta\left(\lambda_{d,1}(t), \lambda_{d,0}(t)\right)$ and $Beta\left(\omega_{d,1}(t), \omega_{d,0}(t)\right)$ for $d = 1, 2, 3$, to numerically estimate the shape of the marginal feature likelihoods.

The density estimates of all d features given one class were inserted into Equation A.2.16. This yields the densities of the stimulus likelihood given class 1 and the corresponding likelihood given class 2. The resulting class-conditional stimulus probability density was then used in Equation A.2.5.

A.2.5 Bayesian model averaging

Let w be the posterior structural belief. Given w, the DISC-LM combines the class predictions as a weighted sum:

$$\hat{p}(c \mid s) = w\,\hat{p}(c \mid s; cci) + (1-w)\,\hat{p}(c \mid s; flex), \tag{A.2.22}$$

where $\hat{p}(c \mid s)$ is the point estimate of the posterior probability that stimulus f belongs to class c, and $flex$ and cci denote that it was generated using flexible feature dependencies or class-conditional feature independence, respectively. Further, $0 \leq w \leq 1$.

The posterior structural belief w is derived by combining a prior structural belief about class-conditional independence, π, with the likelihood of the data given class-conditional independence. The DISC-LM hypothesizes that the structure either follows class-conditional independence or flexible feature dependencies. Let $E_t = e_1, \ldots, e_t$ be data until trial t. Given E_t, the model integrates the prior structural belief with the likelihood:

$$P(\hat{p}(s,c;cci) \mid E_t) = \frac{P(E_t \mid p(s,c;cci)) \cdot \pi}{P(E_t)}, \tag{A.2.23}$$

where π is the prior structural belief, $0 \leq \pi \leq 1$. $p(s,c;cci)$ are the joint prior probabilities of stimuli and classes from the model that assumes class-conditional independence. The denominator $p(E_t)$ normalizes the term and is computed by $P(E_t) = P(E_t \mid p(s,c;cci))\pi + P(E_t \mid p(s,c;flex))(1-\pi)$. We set the denominator to 10^{-30} in cases where $\pi = 0$ or $\pi = 1$, which caused the denominator to be zero and the fraction to be ill defined. We worked with densities that were approximated using Monte Carlo simulation.

The likelihood of the data given class-conditional independence is computed by

$$\hat{p}(E_t \mid p(s,c;cci)) = \prod_i \prod_j p(s,c;cci)^{N_{i,j}(t)} \tag{A.2.24}$$

where i indexes stimuli, j indexes classes, and $N(t)$ denotes how often each combination of stimuli and classes occurred until trial t; formally $N_t = \sum_t N_{s_1,c_1,t}, \ldots, N_{s_8,c_2,t}$.

We log-transformed this calculation to avoid numerical errors:

$$\sum_i \sum_j \exp\Big(\log(P(p(s_i,c_j;cci)))N_{i,j}(t)\Big), \tag{A.2.25}$$

, where the notation is the same as above. To ensure the logarithm was defined we used 10^{-60} if $p(s,c;cci) = 0$.

A.2.6 Simulation results: Conservatism Parameter δ

The following figures (Figures A.2.2 and A.2.3) show simulation results from the DISC-LM where we vary the conservatism parameter δ and the values of the prior structural belief parameter π.

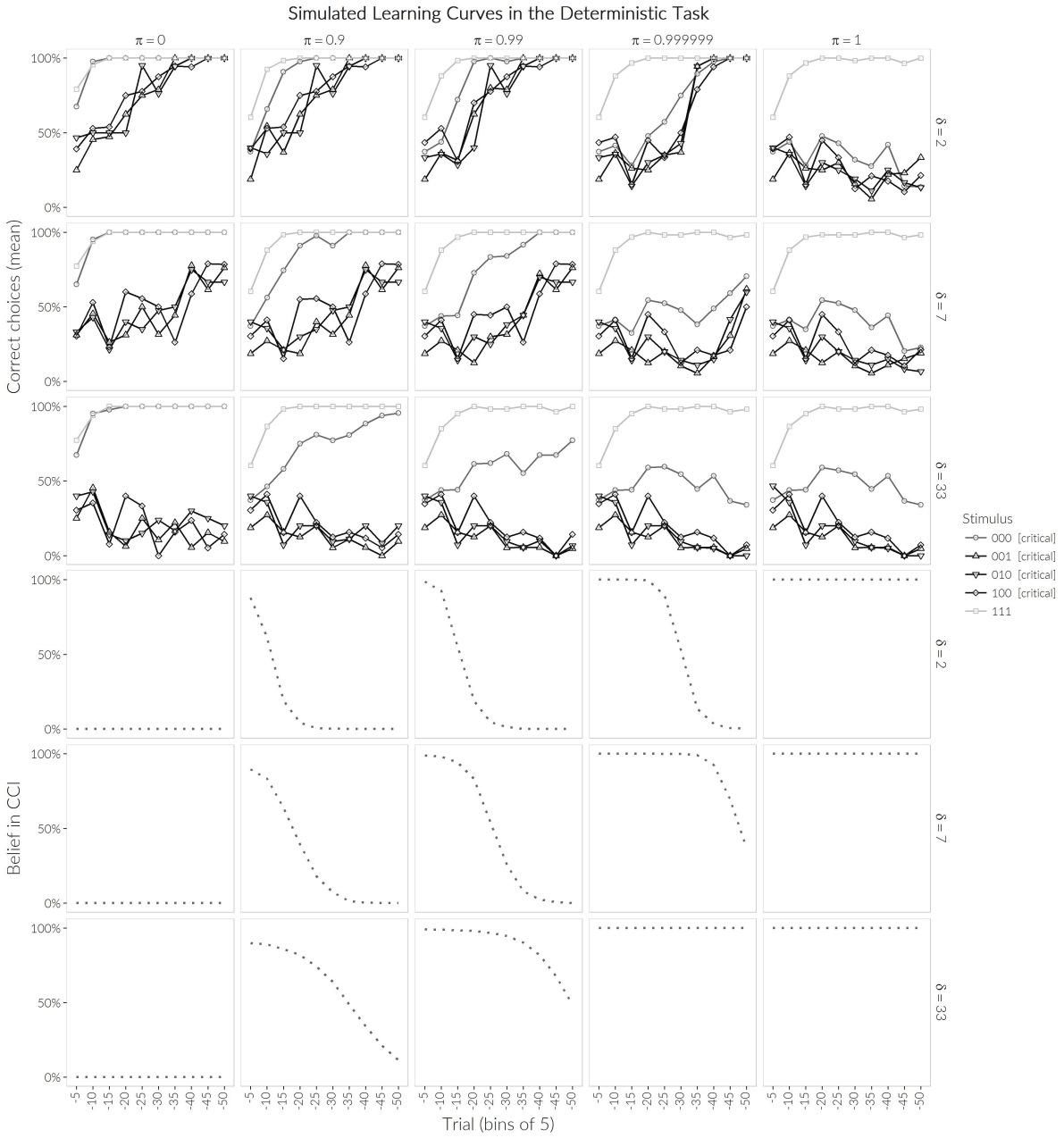

Figure A.2.2. Simulation of the DISC-LM (environment 1). The higher the values of the conservatism parameter δ, the slower learning. Higher values of δ slow down learning for all stimuli, whereas higher values of π affect the critical stimuli but not the uncritical stimulus 111. CCI = class-conditional independence.

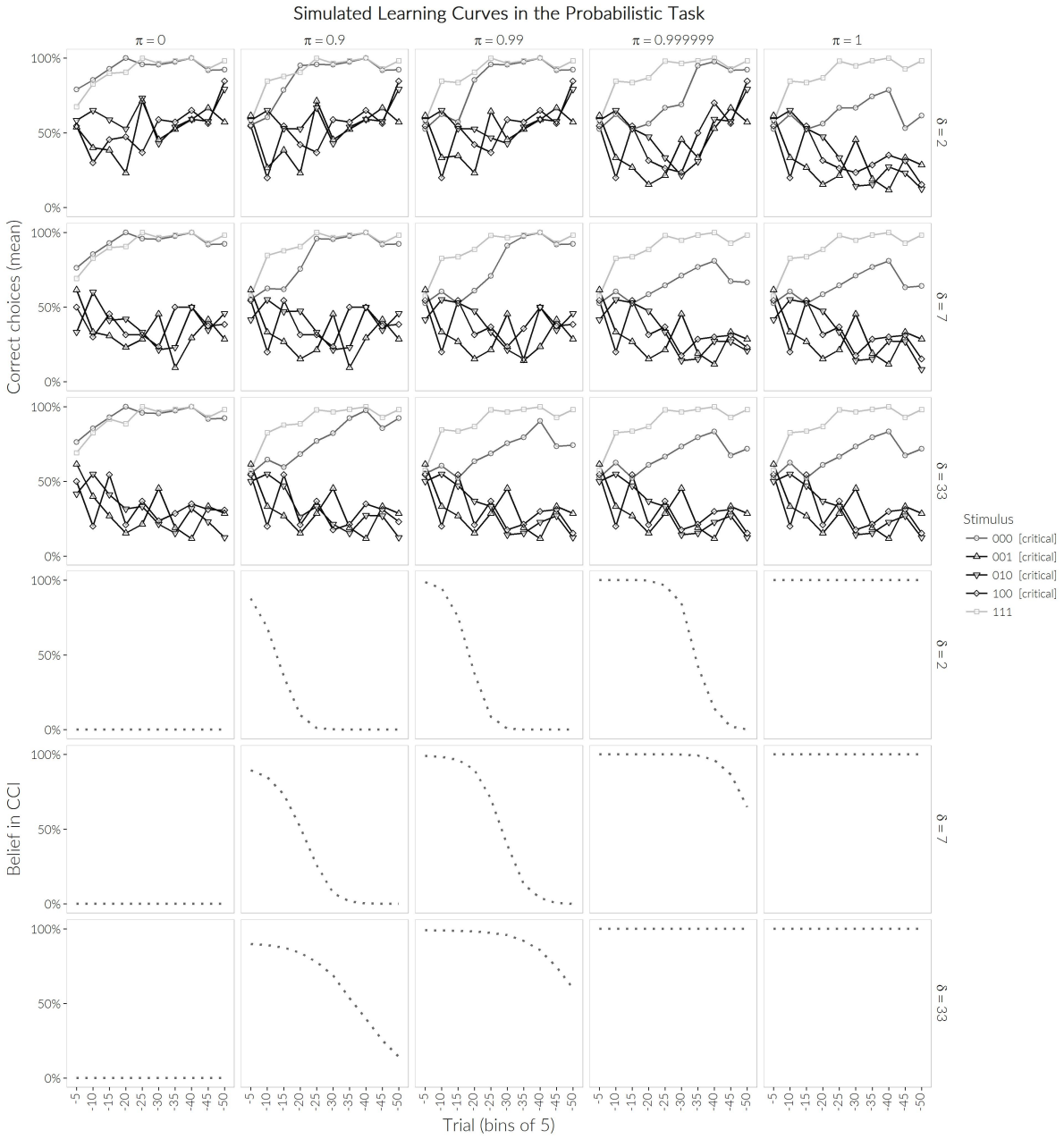

Figure A.2.3. Simulation of the DISC-LM (environment 2). The higher the values of the conservatism parameter δ, the slower learning. Higher values of δ slow down learning for all stimuli, whereas higher values of π affect the critical stimuli but not the uncritical stimulus 111. CCI = class-conditional independence.

A.2.7 Experimental instruction: Feedback during learning

The feedback that participants received every 100 trials was as follows:

How are you doing? If you continue responding like this in the last 200 trials, you will average about x% correct. The optimal strategy achieves about y%.

Mini-FAQ: Q: I've only learned one feature. Is that okay? A: No. More than one feature matters. You must learn all the features to be able to learn to categorize the plankton specimen.

The variable x was the accuracy that would be achieved on average if the participant would respond in the same way as in the most recent 200 trials, and the stimulus configurations would occur exactly according to their average frequencies. The variable y was the maximum achievable average accuracy, if stimuli would occur according to their average frequencies. (Each stimulus was chosen at random according to the theoretical frequencies of occurrence, in each trial in the learning task. Because of this, a participant's actual accuracy is typically not identical to the theoretical accuracy that would be achieved by their pattern of responses to the various stimuli.) Both numbers were rounded to the nearest tenth of a percent. See Tables 3 and 4 for the expected classification accuracies $P(class \mid stimulus)$ in Experiment 1 and Experiment 2, respectively.

A.2.8 Modeling results: Conservatism parameter

The best predicting values of the parameter governing conservatism (uniform slowness in learning), δ, are shown below in Figures A.2.4 and A.2.5 for Experiment 1 and Experiment 2, respectively.

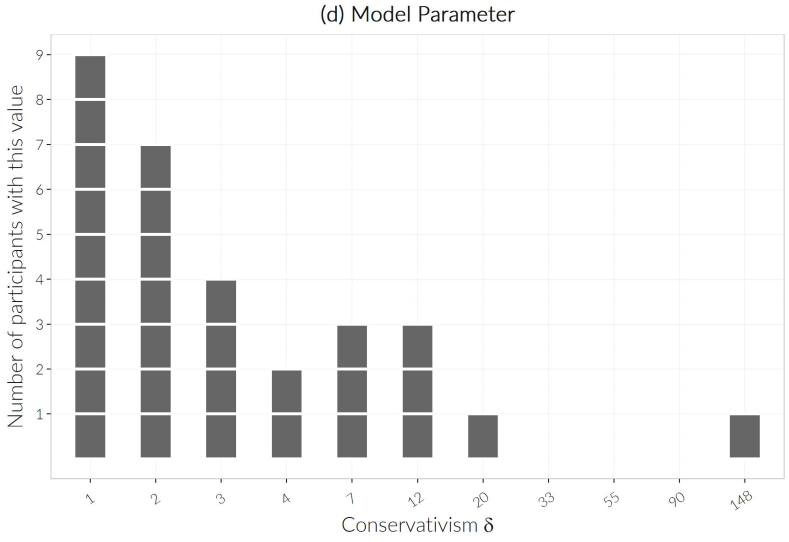

Figure A.2.4. Conservatism parameter from Experiment 1. Note: The fit measure used was mean squared error.

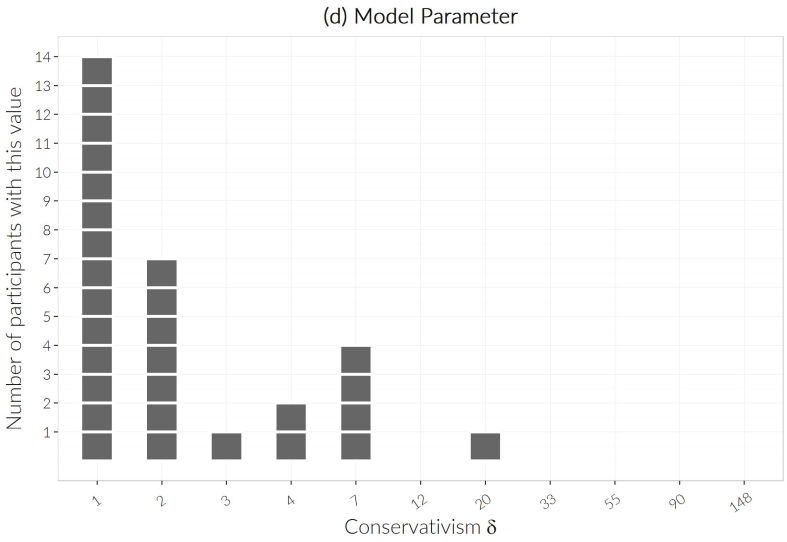

Figure A.2.5. Conservatism parameter from Experiment 2. Note: The fit measure used was mean squared error.

A.3 Jarecki & Wilke (in preparation)

A.3.1 Instructions of Study 1

The instructions for the aspect listing in study 1 read as follows:

> *Under which conditions would you personally be more [less] likely to engage in the described behavior and under which conditions would you personally be less [more] likely to engage in the described behavior? Please list one situational aspect at a time. Enter the first aspect in the box below and, as soon as you are done, hit the "Enter" key to submit.*
>
> If the situation was ... I would be ... (upon clicking this text disappeared)

After the first aspect was entered, the following text appeared:

> *Please continue to list the circumstances that would make you personally more [less] likely or less [more] likely to engage in the described behavior. Please keep going until you cannot think of any more. If you need to pause to consider more aspects of this situation, feel free to do so.*
> *Enter the next situational aspect in the box below and, as soon as you are done, hit the "Enter" key to submit.*

A.3.2 Methods: Details of the Qualitative Data Analysis

We generated the coding manuals from a subset of statements. These statements were chosen identically for each domain. We obtained the statements from the following four participants: two men and two women, one above and one below median age; specifically the person with most responses among the women below, the women above, the men below, and the men above median age. The assistant and JJ coded domains separately. The rason for this is that we wanted a heterogeneous sample, because evidence suggests that women are more risk-averse than men, and older more risk-averse than younger people.

We took four steps in the manual preprocessing of the qualitative data, broadly resolving discrepancies and applying manual changes.

(1) The raters discussed all *coder discrepancies* (statements for which coder 1 and coder 2 disagreed); and resolved them settling on one category.

(2) Thereafter we split *multiple codings* (statements for with both coders assigned multiple codes) into separate statement with separate codes. For example "I would be more likely (...) if their ideas were bad or I had an especially good one (...)" was separated into two codes. Splitting increased the total number of cues for the participant by +1. We preserved the original order of cues. Importantly, we

decided also to split compound statements (e.g., the text "If there were few fans of my team there, *and* my support would make a difference (...)", emphasis added). The reason was that with study 1 we were really only interested in collecting frequencies of individual cue categories, and compound cues were very infrequent (33 of 1598 statements); further we investigated the usage of compound cues in study 2.

(3) We removed *errors* (e.g., two people pressed the enter key before their text, and they wrote the corrected version as the next statement). Removal decremented the total number of a writer's cues by 1.

(4) We discussed *misunderstandings* (statements for which we assumed the writer had misunderstood the question), and removed them. Again, this decremented the writer's total number of cues.

A.3.3 Results: Comparison to Study 2 by Wilke et al.

Table A.3.1 shows the replication results in terms of regression coefficients between the present study and study 2 by Wilke et al., 2014.

Table A.3.1

Comparison of the regression results (risk taking likelihood ~ domain × gender) between the present study and study 2 by Wilke et al. (2014).

	Variable	Present Study			Study 2 by Wilke et al. (2014)			Similar
		Effect	95% *CI*s	*p*	Effect	95% *CI*s	*p*	
Domain	Within-group Competition	0.659	0.302, 1.016	< .001	0.111	-0.096, 0.318	.292	Dir
	Status/Power	-1.607	-1.989, -1.226	< .001	-2.480	-2.701, -2.260	< .001	Yes
	Environmental Exploration	-0.618	-0.989, -0.247	.002	-0.242	-0.454, -0.030	.026	Yes
	Food Acquisition	-0.813	-1.187, -0.439	< .001	-0.854	-1.065, -0.644	< .001	Yes
	Food Selection	0.585	0.219, 0.952	.002	-0.847	-1.059, -0.636	< .001	
	Parental Investment	-0.297	-0.667, 0.074	.117	-0.904	-1.116, -0.692	< .001	Dir
	Kinship	1.839	1.460, 2.217	< .001	1.545	1.325, 1.765	< .001	Yes
	Mate Attraction	0.258	-0.110, 0.625	.171	-1.137	-1.352, -0.921	< .001	
	Mate Retention	-1.056	-1.428, -0.684	< .001	-1.997	-2.212, -1.783	< .001	Yes
Gen.	Female	-0.700	-1.148, -0.253	.003	-0.807	-1.032, -0.583	< .001	Yes
Domain × Gender	Within-group Competition × Female	0.124	-0.381, 0.630	.630	0.651	0.378, 0.923	< .001	Dir
	Status/Power × Female	-0.321	-0.889, 0.246	.268	-0.245	-0.546, 0.057	.112	Yes
	Environmental Exploration × Female	0.421	-0.108, 0.950	.120	0.110	-0.172, 0.392	.445	Yes
	Food Acquisition × Female	0.181	-0.351, 0.713	.505	0.368	0.091, 0.644	.010	Dir
	Food Selection × Female	1.502	0.976, 2.028	< .001	1.715	1.435, 1.994	< .001	Yes
	Parental Investment × Female	0.886	0.355, 1.417	.002	0.972	0.693, 1.251	< .001	Yes
	Kinship × Female	1.273	0.740, 1.806	< .001	1.171	0.884, 1.459	< .001	Yes
	Mate Attraction × Female	-0.460	-0.987, 0.066	.087	0.063	-0.222, 0.349	.663	Yes
	Mate Retention × Female	-0.097	-0.630, 0.437	.724	0.112	-0.173, 0.397	.442	Yes

Note: *Female* denotes a gender dummy with 1 = female. Effects are regression coefficients, Similar summarizes the replication: *Yes = replication* (effects have the same direction and are significant, or both are not significant at $\alpha = .05$), *Dir = direction* (effects have equal direction but only one is significant). The baseline domain (intercept) is between-group competition.

Further, the sociodemographic variables differ in the same manner as outlined in the main text between our sample and study 2 from Wilke et al., 2014; see Table A.3.2. There are significant differences between the present and the past sample regarding age (mean 33 vs. 19 years, respectively), ; relationship status, $t(123) = 13.987, p < .001$, Cohen's $d = 1.785$; number of offspring per person (mean 0.60 vs. 0.057, respectively), $t(205) < .001, p = .486$, Cohen's $d = 0.078$; and the proportion of married participants (54% vs. 22%, respectively), $\chi^2(1) = 38.576, p < .001$, Cohen's $h = 0.676$.

Table A.3.2

Demographics and life history variables in the present study and study 2 by Wilke et al., 2014

Source	Statistic	Female	Age	Married	Children	Siblings	Life Expectancy
Present sudy	Mean	51.7%	33.37	54.20%	0.61	1.56	66.98
	Range		18 – 65		0 – 5	0 – 6	35.5 – 100
Study 2 by Wilke et al.	Mean	58.1%	19.36	42.35%	0.05	1.62	82.83
	Range		17 – 49		0 – 4	0 – 11	34 – 100

A.3.4 Results: Inter rater reliability

Table A.3.3 shows the interrater reliability for categorizing responses in study 1. Two raters coded participants self reports about which aspects of a situation made them more or less likely to engage in risky behaviors across the ten domains. The interrater reliability coefficients are suitable for our data (two raters, multiple nominal categories, and unequal marginal category distributions); for details see Feng, 2014. The measures are Gwet's AC_1 (Gwet, 2008), which performs particularly well when categories are unequally frequent (which is the case in our data).

Table A.3.3

Inter rater reliability for categorizing the situational aspects across the ten domains

Domain	Cohen's $\varkappa$	Krippendorf's α	Gwet's AC_1	Brennan-Prediger's $\varkappa_n$
Between-group Competition	0.61	0.61	0.63	0.62
Within-group Competition	0.72	0.72	0.72	0.72
Status/Power	0.67	0.67	0.68	0.68
Environmental Exploration	0.62	0.62	0.63	0.63
Food Selection	0.76	0.76	0.78	0.78
Food Acquisition	0.67	0.67	0.69	0.69
Parental Investment	0.67	0.67	0.69	0.69
Kinship	0.72	0.72	0.73	0.73
Mate Attraction	0.65	0.66	0.67	0.67
Mate Retention	0.71	0.71	0.72	0.72

Table A.3.4 shows the interrater reliability with respect to the direction of the answer (whether the statement indicated more or less risk-taking) between two raters. For the interrater measures, see Feng, 2014.

Table A.3.4

Inter rater reliability for categorizing the direction across the ten domains

Domain	Cohen's $\varkappa$	Krippendorf's α	Gwet's AC_1	Brennan-Prediger's $\varkappa_n$
Between-group Competition	1.00	1.00	1.00	1.00
Within-group Competition	0.97	0.97	0.98	0.98
Status/Power	0.95	0.95	0.97	0.96
Environmental Exploration	0.97	0.97	0.97	0.97
Food Selection	1.00	1.00	1.00	1.00
Food Acquisition	0.97	0.97	0.98	0.98
Parental Investment	0.98	0.98	0.98	0.98
Kinship	0.95	0.95	0.97	0.96
Mate Attraction	0.98	0.98	0.98	0.98
Mate Retention	0.83	0.83	0.91	0.90

Appendix B

Erklärung über den Eigenanteil

Erklärung über den Eigenanteil an den zur Veröffentlichung vorgesehenen eingereichten wissenschaftlichen Schriften innerhalb meiner Dissertationsschrift gemäß § 6 Abs. 2 Satz 7 der Promotionsordnung der Mathematisch-Naturwissenschaftliche Fakultät II.

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Hilfsmittel benutzt und alle wörtlich oder inhaltlich übernommenen Stellen als solche gekennzeichnet habe.

Hiermit erkläre ich, dass nur die unten genannten Personen an den Studien mitgewirkt haben.

Arbeit

Name:	*Jana Bianca Jarecki*
Institut:	*Mathematisch-Naturwissenschaftliche Fakultät II*
Promotionsfach:	*Psychologie*
Thema:	*Modeling the Decision Making Mind: Does Form follow Function?*

Eingereichte Schriften

1. Jarecki, J. B., Tan J. H., & Jenny, M. A. (under review) What is a cognitive process model? A disambiguation.
2. Jarecki, J., Meder, B., & Nelson, J. D. (2013). The Assumption of Class-Conditional Independence in Category Learning. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Cooperative Minds: Social Interaction and Group Dynamics. Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 2650–2655). Austin, TX: Cognitive Science Society.
3. Jarecki, J. B., Meder, B., & Nelson, J. D. (submitted). The Assumption of Class-Conditional Independence in Category Learning. *meanwhile accepted for publication in Cognitive Science, Dec. 2016*
4. Jarecki, J. B., & Wilke A. (in preparation)